ROADMAP FOR MEASURING LEARNING THROUGH THE LIFECYCLE IN MULTI-TOPIC NATIONAL HOUSEHOLD SURVEYS

2021-09-28

SUMMARY

This report takes stock of a pre-identified set of instruments that measure the abilities of individuals of different ages to meet certain minimum proficiency thresholds on competences, learning and skills ('foundational skills'), which have already been (or can be prepared to be) implemented as part of a household survey. The document includes information on evidence of validity and reliability of these tools, and lessons learned from their administration in different settings. Likewise, the document highlights the potential of and guidelines for the existing tools to be administered through different delivery modalities (one-on-one in person interview, self-administered, and phone-based) and modes (paper-based, computer-based), in the future, including practical proposals to resolve some of the critical implementation trade-offs that emerged at the February 2020 technical workshop.

Table of Contents

1   Introduction

There is a long history of measuring schooling through household interviews. In the United States, individuals have been asked whether they attended school since the census of 1850 (Series 2021). School attendance has been measured in the UNICEF MICS survey since the initial round starting in 1995. However, as made clear in the 2018 World Development report (Filmer et al. 2018), measuring the number of years of schooling is not the same as measuring learning.

A recent World Bank report measuring learning poverty (Azevedo et al. 2021), a metric that combines the dimension of learning with school attendance, found that around 90% of learning poverty can be attributed to learning itself and only 10% to lack of school attendance, meaning that household surveys that only capture schooling may miss significant issues related to learning overall. Additionally, the report highlights significant gaps in the data available to measure learning. In Sub-Saharan Africa, internationally comparable learning assessments are available in only 21 out of 48 countries since the year 2000, making up only 46.8% of the end of primary age (age 10-14) population, and only 13 out of 31 low-income countries globally have an internationally comparable end of primary assessment since 2000. Even less data on learning is available for other age groups, such as for children under age 5 and for adults according to the UN global SDG indicators database.

Multi-topic household surveys that are regularly carried out in most developing countries around the world offer a possible vehicle to collect such data at scale. Countries have committed to report on the multiple education-related indicators under Sustainable Development Goal (SDG) 4, which are meant to go beyond school attendance and measure the extent to which students in that country have reached a minimal level of proficiency in reading and math for their age. This includes at the early childhood level (SDG 4.2.1), grade 2 & 3 (SDG 4.1.1a), end of primary (SDG 4.1.1b), lower secondary (SDG 4.1.1c) levels, and for adults (SDG 4.6.1). Much of this effort has focused on SDG 4.1.1, which measures minimum proficiency in reading and mathematics at different grade levels, and which uses large-scale assessments (either international or national) for data collection, which standards were only approved by the international community in November 2018. Other SDG indicators, such as 4.2.1, which measures early childhood development outcomes, 4.5.1, which measures gender, rural/urban, wealth and other parity in education outcomes, and 4.6.1, which focuses on the literacy and numeracy skills of youth and adults, would require data collection to happen outside of the classroom.

While household surveys cannot substitute for international or national large-scale assessments of in-school students, there are several reasons why such instruments could be complemented by measures of skills or student learning via household surveys. Such household surveys can allow for capturing a representative sample of out-of-school population, be it due to age (i.e. pre-school or post-school) or due to education status (i.e. school-age dropouts). Likewise, the ability to correlate measured skills or competencies with a diverse set of individual-, household-, and community-level variables collected as part of these surveys can provide significant scope for better understanding the factors that could be driving learning outcomes. Furthermore, such surveys allow for applying different administration methods (e.g. caregiver reports for young children and direct assessment for older children and adults) and triangulating direct assessment of an individual's skills or competences with other-reports by household members (e.g., direct assessment of school-age children and caregiver reports) to derive a more comprehensive measure of these skills and competences. In contexts where there is insufficient national capacity or resources to carry out large-scale assessments of in-school students, an assessment of skills or student learning via

household surveys could present an opportunity to build such capacity while temporarily filling an important SDG 4 learning assessment data gap at relatively low cost. Moreover, in fragile and conflict affected contexts or in refugee populations, especially those living in camps, a learning module captured through household surveys might be more effective. Finally, collecting measures of competences, learning, and skills data via household surveys allows for the linkage of data on learning to administrative and service delivery data (on health, education, transport, etc.).

The COVID-19 pandemic has provided an additional reason to focus on household survey data collection, as school closures forced system-level assessments of student learning, which are usually administered in classrooms, to be put on hold. Given the uncertain timeline for the return to in-person schooling, integration of learning assessment into one-on-one in person or phone-based survey administration can address the need for stakeholders to be informed of students' learning status and to measure some of the learning losses in a systematic and rigorous manner. Moreover, it is expected that COVID related school closures will affect school engagement and might increase dropout rates of selected groups of the population. Going forward it will be critical for policy makers to identify the risk factors that might affect the dropout, and one critical question is to what extend past levels of learning, and the learning levels of household, measured either through siblings or parents might affect the school engagement.

In a nutshell, household surveys can be of critical importance to help countries understand the drivers of learning, which can be used to improve the design and targeting of policies to accelerate learning. The complementarity of such efforts with the SDG 4 hinges on the use of a similar benchmark to identify students below a minimum proficiency level, building on the same Global Proficiency Framework (GPF). This is the motivation for this activity.

In this context, the World Bank (WB), the United Nations Educational, Scientific and Cultural Organization (UNESCO) Institute of Statistics (UIS), the UNESCO Institute for Lifelong Learning (UIL), United Nations Children's Fund (UNICEF), International Association for the Evaluation of Educational Achievement (IEA), and the Organisation for Economic Co-operation and Development (OECD) are collaborating on a work program that would enable countries to administer short learning assessments as part of household surveys. The objective of this report is to document the set of instruments that measure the abilities of individuals to meet certain minimum proficiency thresholds on competences, learning and skills ('foundational skills'), which have already been (or can be prepared to be) implemented as part of a household survey. The document includes information on evidence of validity and reliability of these tools, and lessons learned from their administration in different settings. Likewise, the document highlights potential of and guidelines for the existing tools to be administered through different delivery modalities (one-on-one in person interview, self-assessment, and phone-based) and modes (paper-based, computer-based), in the future, including practical proposals to resolve some of the critical implementation trade-offs.

This work will lay out a roadmap that results in a set of modules for use in household surveys to measure skills or student learning. The goal is for each module to have a maximum length of 30 minutes per respondent and to be able to discriminate around relevant proficiency thresholds aligned with the SDGs. The modules produced as a result of this work are meant to be a global public good and will be based on the following principles:

1. **Survey modules will be made available to any organizations conducting household surveys, so long as they abide by licensing rules and guarantee that items remain confidential.** This means that there should be a clear process for any organization conducting

a household survey to include the proposed learning assessment modules in this report. Any organizations following the license rules and with a system in place to ensure the confidentiality of the items should be permitted to use the survey modules. Users of the modules must give appropriate credit to the custodians of the module and indicate if changes were made. Users may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. Because the modules are being made available as a global public good, users are encouraged to respond in turn by making microdata available publicly for use by the international community to the extent possible.

2. **Survey modules should be made available for use free of charge.** This principle does not apply to the cost of administering the survey or of possible advisory services that may accompany conducting such a survey. However, for organizations qualifying for and abiding by the licenses that may be applicable and that guarantee the security of the items, the contents of the survey modules necessary for conducting the survey will be made available free of charge.

3. **Survey modules and the item response scoring should, to the extent possible, not be constrained to any particular survey platform, software platform, or firm.** There should not be terms of use that dictate a particular survey platform or software platform to be used to conduct the survey or item response scoring. For instance, it should be possible to carry out the survey on any number of survey platforms (Survey Solutions, Survey CTO, CS Pro, etc), and after carrying out the survey, it should be possible to score the item responses without necessarily sending off the data to a specialized firm for scoring.

2   Where do we currently have data on learning?

In the following section, a discussion will be given on the current availability of data containing measurement of learning across the lifecycle. Particular attention is given to the availability of data for reporting on the Sustainable Development Goals (SDGs), as it represents an agreed upon set of goals for countries to meet and because the vast majority of countries, through the SDG dialogue, have agreed to report on indicators related to these goals. Coverage gaps across regions, income groups, ages, and years are discussed. Additionally, a discussion is given in how much variation in learning outcomes is explained by factors captured during in-school assessments.

2.1   Gaps in SDG indicators related to education

The SDG goals and targets provide a comprehensive, internationally agreed framework, encompassing indicators reflecting all domains of official statistics: economic, social, environmental and institutional which is relevant to all nations. The targets and indicators for SDG 4, on quality education, are below in Table 1. The indicators cover roughly the ages 2-5 (SDG target 4.2), ages 7-14 (SDG target 4.1), and ages 16+ (SDG target 4.6). Additionally, SDG target 4.5 focuses on equity across all age levels, particularly with regard to gender and the vulnerable groups. Importantly, the indicators go beyond measuring participation in schooling and require the measurement of proficiency in early childhood skills, reading and numeracy in primary and secondary, and adult literacy and numeracy.

Table 1. SDG 4 Indicators

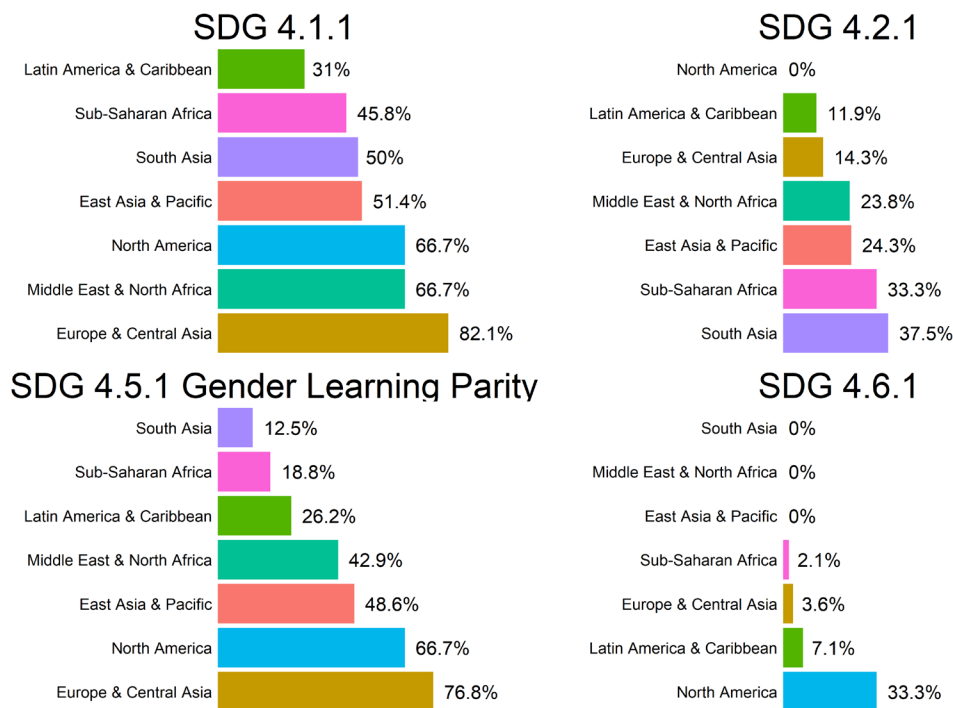| Target	Indicator
4.1 By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes	4.1.1 Proportion of children and young people (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex
	4.1.2 Completion rate (primary education, lower secondary education, upper secondary education)
4.2 By 2030, ensure that all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education	4.2.1 Proportion of children aged 24-59 months who are developmentally on track in health, learning and psychosocial well-being, by sexi
	4.2.2 Participation rate in organized learning (one year before the official primary entry age), by sex
4.3 By 2030, ensure equal access for all women and men to affordable and quality technical, vocational and tertiary education, including university	4.3.1 Participation rate of youth and adults in formal and non-formal education and training in the previous 12 months, by sex
4.4 By 2030, substantially increase the number of youth and adults who have relevant skills, including technical and vocational skills, for employment, decent jobs and entrepreneurship	4.4.1 Proportion of youth and adults with information and communications technology (ICT) skills, by type of skill
4.5 By 2030, eliminate gender disparities in education and ensure equal access to all levels of education and vocational training for the vulnerable, including persons with disabilities, indigenous peoples and children in vulnerable situations	4.5.1 Parity indices (female/male, rural/urban, bottom/top wealth quintile and others such as disability status, indigenous peoples and conflict-affected, as data become available) for all education indicators on this list that can be disaggregated
4.6 By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy	4.6.1 Proportion of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex

Source: UN.

While nearly all countries have agreed to report on the SDG indicator, according to the UN global SDG indicators database major gaps exist in what indicators are available for targets 4.1, 4.2, 4.5,

and 4.6. The UN global SDG indicators database provides access to the data compiled by the UN for the annual Sustainable Development Goals Report tracking progress toward fulfilling the SDGs (Division 2021). Figure 1 and 2 below show the percent of countries in each region and income group that possesses an indicator between 2017 and 2020 for each of SDG goals 4.1.1, 4.2.1, 4.5.1, and 4.6.1. Additionally, Annex 5.2 contains maps of the availability of indicators for each of these goals.

The methodology for this exercise follows Dang and Serajuddin (2020), which summarizes the availability of SDG indicators in the UN SDG database. For each SDG indicator, a country scores "1" if a value can be found in the database for that indicator between 2017 and 2020, and "0" if data is missing for that indicator between 2017 and 2020. The value for the region and income group in Figure 1 and Figure 2 below is the percentage of countries in that region or income group with a value found in the database. The percentages are raw percentages across regions or income groups and are not weighted for population.

Figure 1. SDG Indicator Availability by Region between 2017 and 2020.

SDG 4.1.1

Region	%
Latin America & Caribbean	31%
Sub-Saharan Africa	45.8%
South Asia	50%
East Asia & Pacific	51.4%
North America	66.7%
Middle East & North Africa	66.7%
Europe & Central Asia	82.1%

SDG 4.2.1

Region	%
North America	0%
Latin America & Caribbean	11.9%
Europe & Central Asia	14.3%
Middle East & North Africa	23.8%
East Asia & Pacific	24.3%
Sub-Saharan Africa	33.3%
South Asia	37.5%

SDG 4.5.1 Gender Learning Parity

Region	%
South Asia	12.5%
Sub-Saharan Africa	18.8%
Latin America & Caribbean	26.2%
Middle East & North Africa	42.9%
East Asia & Pacific	48.6%
North America	66.7%
Europe & Central Asia	76.8%

SDG 4.6.1

Region	%
South Asia	0%
Middle East & North Africa	0%
East Asia & Pacific	0%
Sub-Saharan Africa	2.1%
Europe & Central Asia	3.6%
Latin America & Caribbean	7.1%
North America	33.3%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each region with at least one SDG indicator between 2010 and 2020.

8

Figure 2. SDG Indicator Availability by Income Level

SDG 4.1.1

Income Level	Percentage
Low income	40.7%
Lower middle income	52.7%
Upper middle income	59.3%
High income	61.5%

SDG 4.2.1

Income Level	Percentage
High income	0%
Upper middle income	31.5%
Lower middle income	34.5%
Low income	37%

SDG 4.5.1 Gender Learning Parity

Income Level	Percentage
Low income	18.5%
Lower middle income	32.7%
Upper middle income	48.1%
High income	56.4%

SDG 4.6.1

Income Level	Percentage
Lower middle income	0%
Low income	0%
High income	2.6%
Upper middle income	9.3%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each income group with at least one SDG indicator between 2010 and 2020.

For SDG 4.1.1 only 31% of Latin America and Caribbean countries have an indicator available in the UN SDG database, which is the lowest rate across regions. The Middle East and North Africa region has the highest rate of indicator coverage at 66.7%. Only around 41% of low income countries have an indicator for SDG 4.1.1, while around 60% of upper middle income and high income countries have an indicator in the UN database. When breaking down SDG 4.1.1 into the sub-indicators, SDG 4.1.1a (grades 2/3), 4.1.1b (at the end of primary), and 4.1.1c (at the end of lower secondary), coverage rates are generally lower. For instance, while 41% of low income countries have some indicator available in SDG 4.1.1, the same percentage have an indicator for 4.1.1a, but only 30% have an indicator for 4.1.1b, and 0% have an indicator for 4.1.1c. These are available in figures A2.1.1 and A2.1.2 in the appendix.

There are particular struggles regarding SDG 4.2.1 and SDG 4.6.1. These indicators cover learning prior to formal schooling (ages 2-4 for SDG 4.2.1) and toward the end or after formal schooling (ages 16+ for SDG 4.6.1). The data gaps for these two indicators highlight the need for a scale up of data collection for this area, and these age groups make household surveys a natural vehicle.

For reporting on SDG 4.2.1, UNICEF is the custodian agency and data for this indicator appears to largely come from the MICS ECDI instrument. For reporting on SDG 4.6.1, only countries that completed the OECD PIAAC or the World Bank STEP surveys possess data that can be used for reporting for this indicator. Only between 37%, 35% and 32% of low income, lower middle income, and upper middle income countries, respectively, have an indicator for SDG 4.2.1. For 4.6.1, 0% of low income, lower middle income, and9% of upper middle income countries, respectively, have indicators available for SDG 4.6.1.

SDG indicator availability by region and income group when considering windows of 7 years and 15 years are available in the appendix.

2.2   How can this complement in-school assessments?

Another justification for using adding a learning module to a household survey is that the rich set of covariates available in a household survey could shed light on drivers of learning. In a typical in-school assessment, researchers might have access to a set of covariates such as student age, gender, grade level, grade repetition history, basic family structure (such as whether living with mother, father, both), basic family income (such as whether parents working, how many books at home), immigration status, languages spoken at home, and several school level indicators such as class size, school inputs, and some institutional information.

While this information can be important for explaining differences in student outcomes, data collected in household surveys could provide further insights into the drivers of learning. Woessmann (2016), using data from the Programme for International Student Assessment (PISA) 2003, found that the R^2 from a regression of student mathematics outcomes on a set of family background information, school resource information, and institutional information collected as part of the PISA in-school assessment was 0.34, leaving a significant amount of variation in student outcomes to be explained.

In a household survey, other factors explaining student learning could be considered including the learning levels of the parents, as assessed by an adult literacy assessment, food security questions, detailed consumption/expenditure information about the households, health outcomes of the students and parents, and household dwelling conditions such as clean water and sanitation, as well as electricity. A large scale data collection effort collecting this information could yield important insights.

Finally, the data will help to identify the characteristics and risk factors of dropouts, testing the hypothesis on whether the learning levels of the child (and household members) are a significant risk factor for student dropout. This will enable the policies to prevent and address dropout to be more tailored and better targeted, and thus more effective and efficient. In addition, where population census data, administrative records, and geospatial information is available, we would be able to use small area estimation techniques (i.e. poverty mapping methodology) to generate more granular maps and spatially disaggregated data to further diagnose the issues related to dropout and improve targeting.

3   What are our current tools?

As discussed in the previous section there are serious data gaps in measuring learning across the lifecycle. There are particularly large data gaps at the beginning (SDG 4.2.1) and end (SDG 4.6.1) of the age ranges measured by the SDGs. These ages (below 5 and over 16) are ages where household surveys are particularly well positioned, as individuals in these ages are frequently out of school, so in-school assessments are not a good alternative.

In the following section, a discussion will be given on the various instruments that are currently available to fill these data gaps.[1] In some cases, such as the MICS ECDI and FLS modules, the
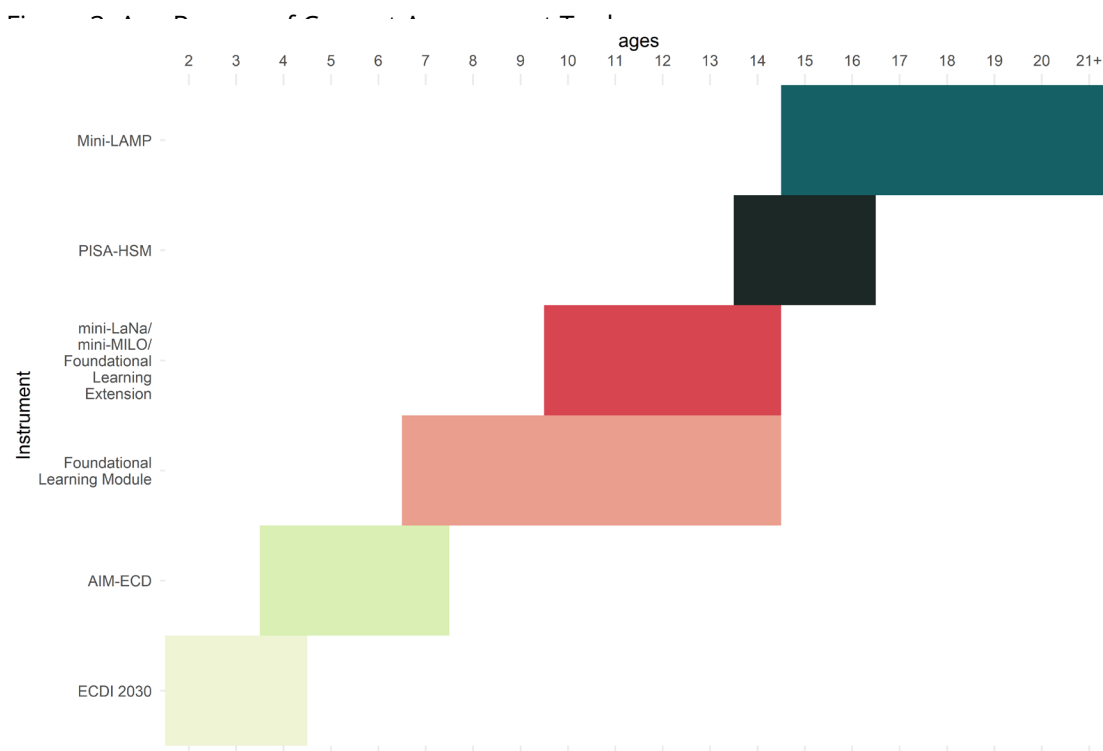
---

[1] There are several other instruments that are potential candidates for measuring learning in a household survey that are not discussed, because they currently lack some of the principles that are discussed in the

instruments have been field tested and relatively little effort will be needed to develop these instruments. In other cases, such as for household modules for measuring student learning at the end of primary, no currently existing assessments exist, and more work is needed. For measuring student learning in secondary school, the PISA-HSM module is in development based on the experience of fielding PISA-D Strand C, and so is in the middle of these two previous stages. In table 3 and figure 3 below, a brief overview of each module is available, along with the age ranges for them. Some of the technical details of the modules are available at the end of the section in table 8.

Table 3. Sample and Age Range of Instruments

Population	SDG	Instrument	Type	Age	Agency
Adults	4.6	Mini-LAMP	Direct Assessment	15+	UNESCO
Secondary	4.1.1c	PISA-HSM	Direct Assessment	14-16	OECD
End of Primary	4.1.1b	mini-LaNa/ mini-MILO/ Foundational Learning Extension	Direct Assessment	10-14	IEA/ ACER/ UIS/ UNICEF
Early Primary	4.1.1a	Foundational Learning Module	Direct Assessment	7-14	UNICEF
Pre-Primary	n/a	AIM-ECD	Caregiver	4-7	WB
ECE for under 5	4.2.1	ECDI 2030	Caregiver	24-59 months	UNICEF

---

introduction: modules fully available to any user, use of items free of charge, no specialized software or company for scoring. They should also be capable of reporting progress on the SDGs. For instance the PIAAC/STEP assessments are not included, as the items cannot be made available, are not free to field, and need to be sent to a specific company to score the items. The EGRA assessment is widely used, but not currently aligned to the Global Proficiency Framework (GPF), which is used for reporting on the SDGs. Also, several citizen led assessments are available, but are also not currently aligned with the GPF. In the future, a version of these assessments could potentially be included.

Figure 3. Age Range of Current Assessment Tools

In the remainder of this section, a discussion of the each survey module will be given with a discussion of the items included, sampling, data collection, and other details of the modules. Because the MICS ECDI and MICS FLS modules are typically conducted in the same survey, these two modules are discussed together, with the remaining modules discussed in the order of the age the modules are targeting. A description of items in the assessment modules can be found in the appendix Annex 5.1.

3.1   MICS

The Multiple Indicator Cluster Surveys (MICS) is a household survey program developed and supported by UNICEF. MICS has been fielded in more than 100 countries since the mid-1990s (Loizillon et al. 2017) and produces data for a range of indicators.

3.1.1   Field Work and Training

A MICS is composed of several questionnaires: a household questionnaire, a questionnaire for women 15-49, a questionnaire for men 15-49, a questionnaire for children under 5 and a questionnaire for children 5-17. These include the Early Childhood Development in the questionnaire for children under 5 and the Foundational Learning Skills (FLS) module that is included in the questionnaire for children 5-17 , both of which will be discussed in greater detail below.

According to the MICS website, fieldwork is usually completed within two to four months.

In terms of sampling, MICS surveys usually target a nationally representative population, with the number of households depending on the country characteristics. Using data included in the World Bank's microdata library, it is possible to examine the historic sample sizes in MICS.

The World Bank's microdata library (https://microdata.worldbank.org/) contains a total of 239 surveys from 99 countries between the years 1995 and 2018. The microdata library documentation contains the number of households that were sampled in each survey. Of the 239 surveys available, 173 of the surveys report the number of households sampled Among these, the number of households sampled ranges from 685 to 300,000, with a median number of households sampled of 6,820.

Figure 4. Number of Households Sampled across 173 MICS Surveys

The 25th percentile of the household sample size across all surveys is 4,703. The median is 6,820 and the 75th percentile is 10,772. Bangladesh 2009 MICS trimmed, due to reported household sample size of 300,000

According to the MICS Survey Plan Template, one randomly selected child in the household is selected for the questionnaire for children 5-17, and one randomly selected for the questionnaire for children under 5.

The recommendation for training of enumerators is for four weeks of training (https://mics.unicef.org/tools#data-collection).

3.1.2   MICS ECDI 2030

The focus of the MICS Early Childhood Development Index (ECDI) is children between the ages of 2 and 5. The module is directed at mothers or primary caregivers, who answer a set of 10 to 20 items (depending on the ECDI module version) on their child's development (Loizillon et al. 2017). The ECDI 2030 module, which is the newest version of the ECDI contains 20 items, while the older version contained 10 items.

Following the adoption of SDG indicator 4.2.1, which measures the proportion of children aged 24-59 months who are developmentally on track in health, learning and psychosocial well-being, UNICEF began a process to update the ECDI items. The resulting module contains 20 questions

directed toward caregivers on their child's development. A typical session is expected to take roughly 3 minutes per child, according to the MICS ECD Instructions for Interviewers.

The MICS ECDI has a high potential for phone adaptation, as the caregiver is asked to give their opinion about the development of the child. For instance, "Can (name) say 10 or more words, like 'mama' or 'ball?'," and "Can (name) recognize all numbers from 1 to 5?"

3.1.3 MICS Foundational Learning Skills Module

The material for this section is drawn from Gochyyev, S. Mizunoya, and M. Cardoso (2019) & Manuel Cardoso (2020). The FLS module focuses on children age 7-14. While the FLS, as it is currently constructed, can provide useful information on the mastery basic skills for 7-14 year olds, it has not been used to report on SDG 4.1.1, which measures minimal proficiency in reading in mathematics as defined by the Global Alliance to Monitor Learning (GAML).

As of September 2021, however, the Bill and Melinda Gates Foundation has provided funding for research to develop a module measuring learning at the end of primary to be incorporated into MICS.[2] This work will strengthen the connection between the MICS foundational learning module and the Global Proficiency Framework (GPF) that has been developed through the GAML.

As for the current version of the Foundational Learning Module:

Reading is measured based on three indicators:
1. % who read 90%+ of words in a story. Story contains 70 words and represents a 2nd grade vocabulary.
2. % who answer 3 out of 3 literal comprehension questions.
3. % who answer 2 out of 2 inferential comprehension questions.

Mathematics is measured based on four indicators:
1. % successfully performing number reading task
2. % successfully performing number discrimination task
3. % successfully performing addition task
4. % successfully performing pattern recognition and completion of missing number task

Gochyyev, S. Mizunoya, and M. Cardoso (2019) uses data from Kenya to calculate the Cronbach's alpha and the inter-rater reliability of the FLS instrument. A total of 48 children were interviewed in a sample of 8 schools in the study.[3] The reading items show Cronbach's alpha measures exceeding 0.8 for all items, while the numeracy items show alpha values of below 0.8, and an overall Cronbach's alpha of 0.76 for the numeracy items. As (Gochyyev, S. Mizunoya, and M. Cardoso 2019), because of

---

[2] https://www.gatesfoundation.org/about/committed-grants/2021/09/inv027808

[3] In psychometrics, internal consistency is an estimate of test reliability. Cronbach's Alpha (and other internal consistency coefficients) can take values between 0 and 1, with tests more consistent as the value of this coefficient approaches to 1. As a general rule of thumb, reliability measures over 0.80 are often considered good and measures over 0.7 considered acceptable. What can be considered good depends on a number of factors including the number of items.

the small number of items in the numeracy assessment, Cronbach's alpha for the numeracy portion should be interpreted with caution.

Table 4. MICS FLS Cronbach Alpha based on Gochyyev, Mizunoya & Cardoso (2019).

Subject	Subtest	Number of Items	Cronbach's alpha
Reading	Oral Reading Accuracy	43	0.91
	Reading Comprehension	5	0.83
	Both subtests' individual items	48	0.92
Numeracy	Number Identification	6	0.54
	Quantitative Comparisons	5	0.61
	Addition	5	0.58
	Missing Number	5	0.62
	Overall numeracy	21	0.76

Source: "Validity and reliability of the MICS foundational learning module". MICS Methodological Papers, No. 9.

Gochyyev, S. Mizunoya, and M. Cardoso (2019) also calculate Cohen's kappa to measure the agreement between two raters of the child's answers in the instrument. These scores are displayed in the table below. The inter-rater reliability for the reading and numeracy items are generally high.[4]

Table 5. MICS FLS Average kappa and PNA based on Gochyyev, Mizunoya & Cardoso (2019).

Subject	Subtest	Average kappa
Reading	Oral Reading Accuracy	0.81

---

[4] A general rule of thumb is that kappa values between 0-0.2 are poor, 0.21-0.4 are fair, 0.41-0.6 are moderate, 0.61-0.8 are good, and values between 0.81-1 are very good (Brennan & Silman (1992)).

Table 5. MICS FLS Average kappa and PNA based on Gochyyev, Mizunoya & Cardoso (2019).

Subject	Subtest	Average kappa
	Reading Comprehension	0.90
Numeracy	Number Identification	0.98
	Quantitative Comparisons	0.96
	Addition	0.93
	Missing Number	0.92

Below is a brief summary of the administration of the FLS module. For more details, consult the MICS Instructions for Enumerators on the MICS website.

FLS module is a direct assessment of children age 7-14, where the enumerator provides verbal guidance to the child during the assessment. Prior to the direct assessment, the interviewers are instructed to receive verbal permission to proceed and to create a rapport with the child. The enumerator instructs the child that the results will not be shared with anyone, and to the extent possible the enumerator is instructed to find a private location for direct assessment, away from caregivers, provided the child wishes to do so. After building rapport, receiving consent, and finding a suitable location for the assessment, the enumerator asks the child a few background questions and then walks the child through each question, giving the child the materials for the exercises.

3.2   AIM-ECD

The World Bank has released two survey modules – a caregiver report and a direct assessment – named the Anchor Items for Measurement of Early Childhood Development (AIM-ECD), for measuring early childhood development for the ages from 48 up to 84 months (4-7 years) old. The AIM-ECD contains items for both a caregiver report of childhood development, as well as a direct assessment. The items for both the caregiver report and the direct assessment cover the developmental domains of early literacy, early numeracy, executive functioning, and social-emotional competencies (Pushparatnam et al. 2021).

The AIM-ECD modules are based on harmonized items from previous measurement efforts including the Measuring Early Learning Quality and Outcomes (MELQO) tool. While the AIM-ECD does not and is not intended to replace the MELQO tool, it does identify a core set of items from caregiver reports and direct assessments to measure children's learning between age 4 and 7, which are made available as a global public good for measuring early childhood outcomes.

While the AIM-ECD module as a whole has not been implemented, at least subsets of the items have been implemented in at least ten countries (Pushparatnam et al. 2021). These include Ethiopia (2017), Kenya (2017), Laos (2016), Lesotho (2017), Madagascar (2016), Mongolia (2015), Nigeria

(2019), Pakistan (2018), Sudan (2015), and Tanzania (2017). Sample sizes in these countries ranged between 160 (Nigeria (2019)) to over 9000 in Laos (2016) in both the caregiver report and direct assessment modules.

Pushparatnam et al. (2021) discusses the psychometric properties of the core set of items for the caregiver report and the direct assessment. The 20 caregiver report items and the 84 direct assessment items were chosen based on the item difficulty, item discrimination, internal consistency, item standardized factor loadings, and relationships to age. The caregiver report contains 5 literacy items and 6 numeracy items, with the other items being executive functioning or socio-emotional items. The direct assessment contains 27 literacy items, 29 numeracy items, and 28 executive functioning or socio-emotional items. (Pushparatnam et al. 2021) anticipate that the direct assessment would take around 30 minutes to complete, with less time for the caregiver report.

Because of the difficulties training enumerators for a multi-topic household survey, the caregiver report version of the AIM-ECD assessment may be advisable, because conducting a direct assessment of a small child requires specialized training of the enumerators. This specialized training includes: how to build a rapport with the child, how to interview the child without discouraging them, how to acquire appropriate consent of the child, and how to demonstrate how the child should perform some of the items. Such specialized training would add to the cost of the survey, so this consideration should be kept in mind.

3.3 GAML and Potential Early Primary and End of Primary Assessments

The Global Alliance to Monitory Learning (GAML), which is led by the UNESCO Institute for Statistics (UIS), is leading a task force to define minimum proficiency levels for the learning targets in the Sustainable Development Goals. For monitoring SDG 4.1.1a (early primary) and SDG 4.1.1b (end of primary), a number of in-school assessments have been mapped to the Global Proficiency Framework (GPF) that has been developed by the GAML including the TIMSS and PIRLS assessment developed by IEA, among other assessments.

However, at this time no household surveys have been used to report for SDG 4.1.1a or 4.1.1b. While this is the case, this section will discuss some of the current assessment tools that could potentially be adapted for a household survey setting or potentially modified to align with the GPF.

3.3.1 Mini-LaNa

The International Association for the Evaluation of Educational Achievement (IEA) has developed the Literacy and Numeracy Assessment for Developing Countries (LaNA) tool for measuring basic literacy and numeracy skills at the end of primary school (4th - 6th grade). The LaNA is meant to be a stepping stone before implementing the TIMSS or PIRLS assessment in developing countries. Using LaNA, countries will be able to report on basic proficiency in literacy and numeracy aligned to the GPF, and, as currently constructed, is administered in a school setting.

According to (Mullis, Ina and Martin, Michael 2015), the items consist of multiple choice items and are scored electronically. The literacy portion consists of a short vocabulary test and reading comprehension items. The LaNA numeracy test contains items on numbers, computation, fractions, and graphs, with items similar to those in the TIMSS assessment. According to (Mullis, Ina and Martin, Michael 2015), the numeracy and reading section each take around 30 minutes, with a combined length of 1 hour.

No version of LaNA has been developed for a household survey, but this could be an area of future work.

3.3.2   MILO

The Global Education Monitoring (GEM) center is collaborating with the Australian Council for Educational Research (ACER) to develop the COVID-19 Monitoring Impacts on Learning Outcomes (MILO) project, which is intended to measure learning outcomes and generate a toolkit for reporting on SDG 4.1.1b (UNESCO and ACER 2021). The project is in the pilot stages, which is expected to begin in 2021 in six countries (Burkina Faso, Burundi, Cote D'Ivoire, Senegal, Kenya, and Zambia). The MILO assessment will be designed to report on the minimum proficiency levels in literacy and mathematics described in (ACER-GEM 2020) and is targeted toward students at the end of primary.

According to (UNESCO and ACER 2021), the MILO assessment will be paper-based and the items are drawn from the UIS Global Item Bank, meaning the results will be aligned with the GPF. The following constructs will be covered: reading comprehension, number and operations, measurement, geometry, statistics and probability, and algebra. The literacy section contains 30 items and the numeracy section contains 30 items.

In order to be incorporated in a household survey, it likely that pilot work will be needed to reduce the number of items.

3.3.3   Foundational Learning Extension

As discussed above, the MICS Foundational Learning Skills module measures learning in children age 7-14. While this age range covers the age range of the SDG 4.1.1 indicators, the FLS module has not at this point been used to report on whether children are meeting the minimum proficiency level in reading and math for these indicators. In the future, the MICS FLS module could be extended to discriminate around the minimum proficiency levels defined for SDG 4.1.1.

3.4   PISA-HSM and PISA-Strand C

While more development is needed to produce a household survey module to measure early primary (SDG 4.1.1a) and end of primary (SDG 4.1.1b) proficiency in reading and mathematics, more development has taken place in developing a household survey module for measuring proficiency at the end of lower secondary (SDG 4.1.1c).

The Programme for International Student Assessment (PISA) has developed two out-of-school assessment for use as part of the PISA for Development (PISA-D) initiative, PISA Strand C and PISA-HSM, which are both designed under the same framework as PISA and PISA-D.[5] The first, PISA Strand C, was created as a household survey module to measure learning in the out-of-school population, 14-16 years old, where respondents answer a set of language and math items that are then scored and scaled using Item Response Theory (IRT), so that a respondents score is on the same scale as PISA. The instrument was fielded to approximately 7,700 youth from 5 countries in 2018. Using the data collected from PISA Strand C, the OECD has developed a shortened version of

---

[5] The PISA-D Out-of-School Assessment Technical Report provides detailed information on the guidelines followed for carrying out the PISA-D out-of-school assessment (PISA-D Strand C), which will be important to follow if adapting this module for another household survey.

the Strand C instrument, named PISA-HSM, which is designed to discriminate around the minimum proficiency threshold for SDG 4.1.1c in a household module that takes less than 30 minutes to administer, compared to around to around 45 minutes for each subject (language or math) in PISA Strand C. The number of items in PISA-HSM is less than 20, and while the assessment would not be as reliable for students away from the minimum proficiency level in reading, would still be reliable around the threshold and would take a significantly less amount of time to conduct.[6]

When considering whether different survey modules can be combined with one another, it is important to discuss the technical requirements needed to implement the modules, so that tablets can be used that meet the minimum technical requirements of all modules. The previously discussed modules do not, as far as the author is aware, require any specialized tablets to implement the modules, which the PISA-D household module may require some specific technical capabilities. The PISA-D Strand C assessment was conducted on custom computer-based platform for tablets. Technical details are available in PISA-D Out-of-School Assessment Technical Report - Chapter 4.

Data collection for the PISA-D Strand C lasted a minimum of 30 days in Senegal to a maximum of 134 days in Guatemala ("Strand C Technical Report - PISA" n.d.). The number of staff receiving assignments ranged from 70 in Paraguay to 220 in Honduras, and the typical hours worked per week by the survey team had a range of 35 in Senegal to 40 or more in Guatemala, Honduras, Panama, and Paraguay.

For field team organization, the recommendation was to have a ratio of 20 interviewers per supervisor. In practice, the ratio of enumerators to supervisors ranged from 4 in Panama to 31.4 in Honduras ("Strand C Technical Report - PISA" n.d.).

The communication between supervisors and enumerators differed across countries. During the main survey data collection supervisors typically communicated with enumerators through phone communications. In Guatemala and Paraguay, daily phone calls were made between enumerators and supervisors and weekly in-person communication was made. In other countries, such as Panama and Senegal, communication made in-person or by phone only as needed ("Strand C Technical Report - PISA" n.d.).

For enumerator training, according to PISA-D Out-of-School Assessment Technical Report - Chapter 7, the countries were, at a minimum, expected to:

• Develop training materials from international versions provided by the contractors.

• Conduct interviewer trainings in person, no more than two weeks prior to the start of data collection.

• Train interviewers in small groups of 15-20.

---

[6] PISA-D Strand C and PISA-HSM surveys are calibrated with Item Response Theory (IRT) on the same scale as PISA. The average and standard deviation of the base year were set to 500 and 100, respectively. The standard errors of the scores around the minimum proficiency threshold of PISA-D Strand C and PISA-HSM are ranged from 20 to 35 points, depending on booklets.

- Assemble training staff to include a knowledgeable lead trainer, assistant(s) and technical support staff.

- Conduct in-person interviewer trainings which must include General Interviewing Techniques as well as the project specific training on the concepts, instruments and procedures related to PISA-D.

- Provide opportunity for supervisory staff to observe and evaluate trainee performance.

- Train interviewers in each language in which they will be conducting the interview and assessment, if there are multiple languages.

Table 6. PISA-D Strand C Training Time Length by Component

Interviewer Training Topic	Estimated Hours
General Interviewing Techniques	3.50
Introduction	0.30
Preparing for the field	0.30
Introduction to the Tablet and the Case Management System	0.50
Locating households/respondents	1.00
Screener interactive, if applicable	2.50
Referral cases	0.50
Youth Interview administration	3.00
Disposition codes	0.75
Exercise Administration	0.75
Administrative procedures	0.50
Quality Control	0.25
Gaining respondent cooperation	1.50
Putting it all together	1.00
Scripted practice interview (role play)	2.25
Live respondent practice	2.00
Total hours for countries with list samples	20.60

Source: OECD/PISA.

Figure 5 below show the number of children interviewed in five countries in which the PISA-D Strand C module was implemented. The sample size ranges from around 1,000 to more than 2,000 students per country.

Figure 5. PISA Strand C Sample Size by Country

3.5 UIS Mini-LAMP

The mini-LAMP (Literacy Assessment and Monitoring Programme) was designed to measure rates of adult literacy and numeracy skills. It followed the UIS LAMP assessment and seeks to improve upon the LAMP assessment by reducing the operational, technical, and financial burden of the LAMP assessment (UIS 2018).

The following skills are measured in the LAMP tool: reading of continuous text, reading of non-continuous texts, and numeracy skills. The LAMP tool was field tested in Afghanistan, El Salvador, Jordan, Lao PDR, Mongolia, Morocco, Niger, Paraguay, Palestine, and Vietnam. The Mini-Lamp contains a subset of the items used in the LAMP tool, containing 15 items measuring the same skills. The assessment is self-administered, following the enumerator providing a set of instructions and collecting background information (UIS 2018). Papers, writing utensils, and calculators are required tools for the Mini-LAMP.

On implementation, while the mini-LAMP has not been implemented, the mini-LAMP team has produced a set of detailed planning documents with recommendations for implementation, including the core and background questions, planning tools, translation guidelines, interview guidelines, data analysis and scoring guidelines. The recommended minimum sample size for Mini-LAMP is such that the survey results in 1500 completed interviews. In order to score the responses, UIL has created a scoring application that can be run on a desktop computer using several common

internet browsers (Microsoft Edge, Chrome, Firefox, Safari, or Opera) or hosted on server. Survey responses to the Mini-LAMP test can be uploaded using this tool and scored, generating plausible values of proficiency on the LAMP scale. Finally, UNESCO has created estimates for the costs of computer based (CAPI) and paper and pencil based (PAPI) implementations of the Mini-LAMP in low-income countries. These estimates are taken from (Montoya, Silvia 2020).

Table 7. Estimated costs of computer based (CAPI) and paper based (PAPI) implementation of Mini-LAMP.

	Task	Sample size = 1,500	Sample size = 3,000	Sample size = 4,500
CAPI	Administering Assessment	$119,350	$198,450	$282,350
	Data Production & Reporting	$38,200	$38,200	$38,200
	Total Cost	$157,550	$236,650	$320,550
PAPI	Administering Assessment	$215,808	$383,680	
	Data Production & Reporting	$37,200	$37,200	
	Total Cost	$252,528	$420,880	

Reproduced from Montoya (2020).

3.6 Summary of Instruments

Table 8 below briefly summarizes a few of the technical details for the instruments discussed in this section. In some cases, such as for the MILO and mini-LaNA instruments, which are under development, details are yet to be determined. The instruments in the table could form the basis for a household survey to measure learning across the lifecycle, but some harmonization would be needed for instance on the sample size, and details would need to be worked out on how to administer such a household survey, potentially with multiple modules being administered to a single household.

For instance, say a household were chosen with one adult respondent and four children ages 3, 5, 10, and 15. In this case, the household would be eligible for the MICS-ECDI2030 instrument (for the 3 year old), the AIM-ECD assessment (for the 5 year old), an end of primary assessment (for the 10 year old), the PISA-HSM assessment (for the 15 year old), and the Mini-LAMP assessment (for the adult respondent). Such a household interview would be expected to take at least 2 hours according to the estimates in Table 8, which may impose a significant burden on the household. It would also require an enumerator or team of enumerators trained for a wide range of assessments. An analysis of how to administer these instruments in tandem is therefore required. In the next section

such a discussion is given, along with a discussion of other details on the implementation of such a survey.

Table 8. Key Technical Details of Instruments

Module	Respondent	Method of Administration	Sample Size Range	Number of Items	Interview duration
MICS-ECDI2030	caregiver	face-to-face	typical range 4,000-11,000 households	20	Under 10 min
AIM-ECD caregiver report	caregiver	face-to-face	range from 160 to more than 9000 children	20	<30 min
AIM-ECD direct assessment	child	face-to-face	range from 160 to more than 9000 children	84	30 min
MICS-FLS	child	face-to-face	typical range 4,000-11,000 households	15	15 min
MILO	child		To be determined	To be determined	unknown
Mini-LaNA	child		To be determined	To be determined	unknown
PISA Strand C	child	self administered	1,000 - 2,100 children	35 Math, 22 Reading, 50 Reading Components (each respondent is given about 40 items in total)	45 minutes
PISA-HSM	child	self administered	To be determined	6 math, 11 reading	30 minutes

Table 8. Key Technical Details of Instruments

Module	Respondent	Method of Administration	Sample Size Range	Number of Items	Interview duration
mini-LAMP	Adult respondent	self administered	at least 1500 interviews	15	30 minutes

4   What is the roadmap forward?

As discussed in the previous sections, substantial data gaps exist across the lifecycle in measuring learning. A shown in table 3 and figure 3, in theory the modules listed above should be sufficient to fill those gaps, but substantial work remains to be done to produce a single survey measuring learning in all of these age ranges. A number of issues need to be resolved including: harmonizing the training of enumerators, the sample design of such a survey, protocols for translation, technical requirements for a potential CAPI survey, scoring, protecting the confidentiality of items, and adding or refreshing items. The information in the previous section can provide some insights into these issues, but future pilot studies will be needed to fully resolve them.

4.1 Implementation Tradeoffs and Considerations

On the 3rd and 4th of February, 2020, experts on learning assessments and from major international survey programs gathered at the World Bank's Headquarters in Washington DC to discuss challenges and opportunities to address gaps in the availability of learning data in low- and middle-income countries. The workshop, entitled "Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries", was convened by the Learning Assessment Platform (LeAP) team, the Skills Thematic Group in the Education Global Practice, and the Living Standards Measurement Study (LSMS) team in the Development Data Group at the World Bank and UNESCO Institute for Statistics (UIS). The meeting brought together a mix of experts and institutions active in the collection of learning and skills assessment via different modalities such as, besides the World Bank and UNESCO UIS, UNICEF, World Bank, the Demographic and Health Surveys (DHS), OECD, Westat, ETS, national agencies and citizen-led assessment organizations from Mexico, Tanzania, and the West Bank.

Several implementation trade-offs were discussed in that workshop, including on survey administration, including on paper based (PAPI) or computer based (CAPI) administration of the surveys, whether to use adaptive testing, and how to localize the modules to be context and culturally relevant, available in local languages, age ranges, ethical considerations, privacy and protection, and enumerator training. These are discussed below with a discussion of future pilot studies to help resolve these issues.

Ethical & Privacy Considerations

Because any assessment of learning involves human subjects, ethical and privacy considerations are critical. This is especially true as the research subjects will often be children. In general verbal

consent should be provided by the child's caretaker and the children/respondent themselves before proceeding. Additionally, the respondents should be made aware that they can stop the interview/assessment at any time with no consequences. To maintain the privacy of the individual's responses, the interview should be done in a location where no other individual is able to hear or see the responses, to the extent possible. For interview with children, a parent or adult known to the child should be visible to the child, but unable to overhear the answers (Mizunoya 2020). In many cases, any organization implementing any of the household survey modules will likely need to seek approval from the national statistics office, an Intuitional Review Board (IRB), and/or a national ethics board before implementing the survey to ensure that procedures are in place to gain consent and maintain the privacy of the respondents.

After data is collected, personally identifiable information (PII) should be separated from the rest of the data as soon as possible and placed in a secure location, or else the data should be stored in an encrypted location to ensure that confidential data is not unintentionally disclosed (Bjarkefur, K., Cardoso de Andrade, L., Daniels, B., & Jones, M. R. (2021))

Background Questionnaires

An area that has not been discussed so far is the key demographic information that will need to be collected for each participant. Many of the modules, for instance the MICS FLS modules, include a set of background questions that are posed to the child or caregiver. In the case of the MICS FLS, a set of background questions on the child such as the child's date of birth, school attendance history, health information, labor force participation of child/caregivers, and survey information such as the location of interview are recorded. Such information would likely be collected for each respondent of the learning modules.

Enumerator Training

Enumerator training is an important part of conducting a household survey, and because enumerators would be expected to assess individuals across a range of ages, the training could be complex and demanding. In particular, this will be the case if the AIM-ECD direct assessment of 4-6 year olds is administered, which involves interview of young children. There may be slight differences in training recommendations across surveys, that will need to be addressed. For instance, according to the MICS training manual, it is advised that a relatively small group of enumerators are trained (less than 100) to ensure uniform application of the survey. The PISA-D Strand C (PISA-HSM) training information recommends enumerators be trained in groups of 15-20 however, and finding the optimal number of enumerators in each training session will require some study. Additional considerations are that PISA-D Strand C (PISA-HSM) recommends that the training takes place less than two weeks before data collection and that the training be done in the local language in which they will be conducting the interview.

Also, the PISA Strand C manual provides an estimated time length of the training of 20.6 hours (see table in section 3.4). At least 8 of these hours would be devoted to the items in the PISA assessment and therefore could not be combined with the training for other modules. Assuming at least 8 hours are needed for training specific to each module, then conducting the training for six modules will require at least 6 days of module specific training, plus more time for other training on field operations, tablet training, and other modules that may be included in the survey, such as a

household background questionnaire. According to the MICS training documentation, the recommended training length is around four weeks.[7]

Sampling

On sampling, the sample sizes for the modules tend to range from a few hundred, in the case of some instances of the AIM-ECD items, to several thousand in the case of the MICS. The minimum recommended sample size for the mini-LAMP is 1500 respondents, the PISA Strand C saw between 1,000 and 2,000 respondents interviewed, while the MICS included a median of 6,800 respondents. Because the MICS includes interviews of individuals across difference age ranges (under 5, 7-14, as well as adults), it may be a good guide for the sample size needed for a combined survey containing the six potential modules discussed in this paper. The 25th percentile MICS included around 4,700 interviews. The median was around 6,800, and the 75th percentile was over 10,000 interviews.

Implementation

One implementation tradeoff is between a tablet based (CAPI) or paper based (PAPI) assessment. The main advantages of most of the CAPI applications are: (i) the creation of complete log files, recording all actions carried out during the interview (i.e. keystrokes, response time, back-ups, etc.), (ii) the validation and test for outliers or unusual values, which allows for the immediate detection and correction of errors, and (iii) the correction of an imperfect sampling frame by using digital maps and ensuring sufficient coverage in the second phase. There are methodological and logistical advantages of using CAPI, such as durability, functionality, and savings in printing and delivering. Potential disadvantages include the additional costs of programming, problems with servers and data storage off-line, and lack of technical know-how (Wild 2020). Additionally, for assessments that have previously been conducted using a paper based assessment, switching to a tablet based assessment could introduce mode effects which jeopardizes comparability. Mode effects are such that an individual's test score may differ because of the mode of the assessment (tablet or paper) (i.e. because of unfamiliarity with tablets).

There are also pros and cons with adaptive testing. Currently none of the modules are set up with a computer adaptive version, but this could be a development for the future. Adaptive tests allow tailoring a test for the ability level of examinees (Tamassia 2020). Adaptive algorithms optimize the delivery of tests items to match the characteristics of individuals, thereby allowing the test to provide more reliable information about skills in a relatively shorter time. Depending on the design, these algorithms can tailor both the difficulty and number of items to the characteristics of a respondent, thus maximizing precision of measurement across a larger range of abilities. While adaptive tests provide an opportunity to accumulate greater accuracy in making the decision, they require more sophisticated methodologies for analysis and larger sample sizes. Furthermore, adaptive algorithms assume that automatically scoreable items are used, so that items that cannot be automatically scored are not usable. The benefits of adaptive testing include (i) Improved measurement without increasing testing time; (ii) At the individual level, benefits in terms of time

---

[7] The MICS does include a number of modules that may not be included in a general household survey, such as on Anthropometry and Water Quality testing.

and efficiency of measurement by ensuring students answers items around their levels of ability, and (iii) At the country level, it better targets the average ability of test-takers.

For potential phone adaptations of the household survey, (Dabalen et al. 2016) mention that a rule of thumb, mobile phone surveys typically last around 15-30 minutes and contain around 20 questions, although the authors note that this time length is simply a rule of thumb and could be studied further. A concern with relying on phone based interviews of caregivers or children is response fatigue. (Abay et al. 2021) highlights significant impacts of response fatigue on results of dietary diversity in Ethiopia. In an experiment, authors find that delaying the arrival of the dietary diversity module by 15 minutes in a phone based survey leads to an 8-17 percent underestimation in the respondent's dietary diversity score. Such considerations would be important to factor in when designing any phone based survey of learning. Additionally, several of the modules have low potential for phone adaptation (see the appendix), besides the caregiver report modules, so more work may be needed to create custom modules for other age ranges.

Translation & Localization

Translation and localization are additional concepts to consider in producing relevant information from a harmonized survey. Children and adults should be assessed in a language that is familiar to them to the extent possible, and asked about concepts that are relevant to the local culture. In general, the following will need to be translated or adapted: the cognitive items, any background questions, the instructions, child stimulus items, and any manuals for enumerators to conduct the survey. Several of the modules, for instance the MICS modules or the World Bank AIM-ECD modules may have translations that have already been implemented in the field. It may be advisable to reach out to those organizations for translations that have been field tested.

Additionally, it would be important to make sure items are relevant to the local context. The protocol for the MICS Foundational Learning Skills module is to choose words appearing in local 2nd grade language textbooks and reading materials (Mizunoya 2020). There is also a collaboration with the national survey team or ministry of education specialists to ensure the items are relevant.

Other Considerations

On other considerations, it will be important to create a harmonized Memorandum of Understanding (MOU) with licensing requirements for governments, survey firms, and other partners to sign in order to use the modules. Each custodian agency of the modules (UNICEF, UIL, UIS, World Bank, OECD, etc.) has a separate MOU currently, but work to harmonize these MOUs would be useful and will require a discussion with each organization.

Additionally, more discussion is needed regarding adding or replacing items in each module. Ensuring the confidentiality and relevance of items is important for the quality of surveys going forward. Clear protocols should be given to users of the modules on how to include new items, while retaining comparability with other surveys and the process for incorporating new items.

4.1   Future Pilot Study

The next steps to produce a working set of modules are a set of pilot studies. This includes piloting the modules that are still under development (mini-LaNA, mini-MILO, mini-LAMP), but also piloting combining the modules into a single assessment. This will include streamlining the administration of the modules to keep respondent burden as small as possible. This may mean randomly assigning

modules to households, so that no household receives more than two modules. It will also include piloting enumerator training to keep costs down.

Based on the analysis in section 2, it is possible to define some parameters of what a future survey may look like, which can provide some sense of the length of time of the survey and the costs. It is assumed that in this future survey, a maximum of one respondent per module per household will be interviewed, and that each module takes 30 minutes to complete.

Using data from 15 countries (i.e. Vietnam, Kyrgyz Republic, Romania, Brazil, Colombia, Peru, Egypt, Morocco, Bangladesh, Pakistan, Angola, Ethiopia, Kenya, Nigeria, and Uganda) from 6 regions, it is possible to calculate the average number of modules that can be given per household. The modules correspond to the following age ranges:

- ECE under 5 (SDG 4.2.1): 2-5

- Pre-primary (na): 4-7

- Early primary (SDG 4.1.1a): 7-14

- End of primary (SDG 4.1.1b): 10-14

- Lower secondary (SDG 4.1.1c): 14-16

- Adult (SDG 4.6): 15+

Population	SDGs	Instrument	Type	Original Age	Agency	Field Work Age (Pilot)
Adults	4.6	Mini-LAMP	Direct Assessment	15+	UIL	15-65
Secondary	4.1.1c	new PISA-D	Direct Assessment	4-16	OECD	14-19
End of Primary	4.1.1b	mini-LaNa / mini-MILO / Foundational Learning Extension	Direct Assessment	10-14	IEA / ACCER / UNICEF	10-14
Early Primary	4.1.1a	Foundational Learning Module	Direct Assessment	7-14	UNICEF	7-14
pre-primary	na	Pre-Primary Module	Caregiver	5-7	WB	5-7
ECE for under 5	4.2.1	new ECDI	Caregiver	2-5	UNICEF	2-5

Using data from these 15 countries, the table below shows the number of households interviewed in each country along with the average number of modules that could be given per household in columns 4 and 5. In columns 6 through 8, the number of households that are needed to produce a

minimum of 200, 300, and 600 interviews per module are shown.[8] In columns 9 through 11, the average length of a household interview is shown, assuming each module takes 40 minutes.[9]

The choice of the minimum number of module implementation can have significant effects of the number of household visits in the absence of a more sophisticated sample design. To ensure a minimum 600 interviews per module, while preserving representation of the household composition, we need to visit on average 2,700 households. The average number of modules per household 2.8 with important regional variations, and the average additional time requirement to implement those modules is approximately 111 minutes taking to the consideration the combination of direct assessment and care giver report.

Number of households, average number of modules per household, and average time of survey across 16 countries.

Region	Country	Year	Number of households	Average number of modules	Households needed			Average time length		
					200	300	600	200	300	600
East Asia and Pacific	Vietnam	2018	8,826.0	2.4	1,154.0	1,691.0	3,178.0	95.9	95.6	96.8
Europe and Central Asia	Kyrgyz Republic	2019	18,478.0	2.7	732.0	1,093.0	2,372.0	112.0	112.2	109.6
Europe and Central Asia	Romania	2019	5,015.0	1.5	5,015.0	5,015.0	5,015.0	59.3	59.3	59.3

---

[8] 200, 300, and 600 interviews are seen as minimums needed to report nationally representative averages for countries. We assume that at least 200-300 interviews are needed to be able to report a single statistic for a country, and 600 may be needed to produce breakdowns, for instance by male/female or urban/rural.

[9] It is assumed that the assessments can be completed in around 30 minutes with instructions taking another 10 minutes for a total of 40 minutes.

Number of households, average number of modules per household, and average time of survey across 16 countries.

Region	Country	Year	Number of households	Average number of modules	Households needed			Average time length		
					200	300	600	200	300	600
Latin America and Caribbean	Brazil	2019	137,747.0	2.0	1,575.0	2,491.0	5,010.0	82.2	81.3	81.0
Latin America and Caribbean	Colombia	2019	217,061.0	2.2	1,808.0	2,606.0	4,977.0	85.9	86.7	86.8
Latin America and Caribbean	Peru	2019	30,599.0	2.6	1,024.0	1,524.0	3,062.0	102.2	102.2	102.7
Middle East and North Africa	Egypt, Arab Rep	2017	11,731.0	2.9	807.0	1,166.0	2,200.0	115.8	115.9	116.5
Middle East and North Africa	Morocco	2013	15,628.0	2.8	896.0	1,360.0	2,557.0	110.0	109.9	111.7
South Asia	Bangladesh	2016	45,380.0	2.9	889.0	1,304.0	2,596.0	113.6	114.6	114.3
South Asia	Pakistan	2018	24,699.0	3.5	442.0	650.0	1,336.0	140.5	140.1	141.9

Number of households, average number of modules per household, and average time of survey across 16 countries.

					Households needed			Average time length		
Region	Country	Year	Number of households	Average number of modules	200	300	600	200	300	600
Sub-Saharan Africa	Angola	2018	11,721.0	3.5	487.0	715.0	1,447.0	140.5	140.6	137.9
Sub-Saharan Africa	Ethiopia	2015	29,741.0	2.9	709.0	1,055.0	2,102.0	119.5	118.7	118.0
Sub-Saharan Africa	Kenya	2015	21,037.0	3.1	615.0	922.0	1,759.0	124.7	125.6	127.3
Sub-Saharan Africa	Nigeria	2018	20,850.0	3.3	467.0	705.0	1,477.0	136.8	136.3	133.9
Sub-Saharan Africa	Uganda	2016	15,270.0	3.2	571.0	876.0	1,659.0	127.6	127.1	127.7
Total			40,918.9	2.8	1,146.1	1,544.9	2,716.5	111.1	111.1	111.0

The table below shows the number of household interviewed in each country along with the average number of modules that could be given per household in columns 4 and 5. In columns 6 through 8, the number of households that are needed to produce 200, 300, and 600 interviews per module are shown.^[200, 300, and 600 interviews are seen as minimums needed to report nationally representative averages for countries. We assume that at least 200-300 interviews are needed to be able to report a single statistic for a country, and 600 may be needed to produce breakdowns, for instance by male/female or urban/rural.] In columns 9 through 11, the average length of a household interview is shown, assuming each module takes 40 minutes including instructions.

4.2   Discussion

Another area of future work will be a research agenda that can build on these household surveys. The ability to correlate measured skills or competencies with a diverse set of individual-, household-, and community-level variables collected as part of these surveys can provide significant scope for better understanding the factors that could be driving learning outcomes. Furthermore, such surveys allow for applying different administration methods (e.g. caregiver reports for young children and direct assessment for older children and adults) and triangulating direct assessment of an individual's skills or competences with other-reports by household members (e.g., direct assessment of school-age children and caregiver reports) to derive a more comprehensive measure of these skills and competences.

5   Final Remarks and Next Steps

A next step in this process will be to implement the pilot studies discussed in the previous section. This includes planned pilots. For 2-5 pilot countries, it is expected to take 18-24 months, as the survey usually runs for several months, often a full year. It will also depend on the timing of planned surveys, as the learning assessments will have to be run in parallel or sequentially to those surveys.

A rough ballpark range to integrate the learning measurement into the national household survey will be $500k-1m per country, depending on country choice and local costs. For instance, for Uganda the additional cost would be in the ballpark of $700-750k, including analytical support and report writing, data preparation, documentation and dissemination, etc. Notice that this cost builds on the existing investment that the country is already making in running their household surveys. These costs in the pilot are not expected to cover the full cost of running a multitopic household survey, which on average is estimated to cost approximately 2 million dollars.

It is also estimated that the pilot costs will be higher than the scale-up phase, since at this stage we will also be covering the development of the tools and creation and documentation of protocols that will need to be used at the scale-up phase.

Below is a list of four countries with upcoming household surveys as part of the LSMS, which could be potential pilot countries:

- **Uganda.** Data collection should start around September 2022. The main survey is funded, but funding will be needed for the add-on on learning modules.

- **Cambodia.** The main LSMS survey is funded, but funding will be needed for the add-on on learning modules.

- **Malawi**. LSMS are planning a time use study with a sample size of 1,500. With co-funding this could be expanded.

- **El Salvador**. A survey is now being planned for two regions in rural areas. With co-funding, it could be possible to expand the scope to add the learning modules.

Following these pilot studies, a guidance document will be produced laying out lessons learned from the pilots including an implementation guide. The next phase will be the scale up to incorporate the modules and begin reporting on SDG indicators.

This work has laid out a roadmap to produce a set of modules for use in household surveys to measure skills or student learning. The document includes information on evidence of validity and reliability of these tools, and lessons learned from their administration in different settings. Likewise, the document highlights potential of and guidelines for the existing tools to be administered through different delivery modalities.

A   Annex

A.1   Item Characteristics

The following tables show for each module the basic item descriptions, response format, whether each item is phone adaptable, and the respondent.

A.1.1   MICS ECDI2030

Table A1.1 MICS-ECDI2030 Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Can (name) walk on an uneven surface, for example, a bumpy or steep road, without falling?	Yes/No	Easy	caregiver
Can (name) jump up with both feet leaving the ground?	Yes/No	Easy	caregiver
Can (name) dress (him/herself), that is, put on pants and a shirt, without help?	Yes/No	Easy	caregiver
Can (name) fasten and unfasten buttons without help?	Yes/No	Easy	caregiver
Can (name) say 10 or more words, like 'mama' or 'ball'?	Yes/No	Easy	caregiver
Can (name) speak using sentences of 3 or more words that go together, for example, "I want water" or "The house is big"?	Yes/No	Easy	caregiver
Can (name) speak using sentences of 5 or more words that go together, for example, "The house is very big"?	Yes/No	Easy	caregiver
Can (name) correctly use any of the words 'I,' 'you,' 'she,' or 'he,' for example, "I want water" or "He eats rice"?	Yes/No	Easy	caregiver
If you show (name) an object (he/she) knows well, such as a cup or animal, can (he/she) consistently name it? By consistently we mean that (he/she) uses the same word to refer to the same object, even if the word used is not fully correct.	Yes/No	Easy	caregiver

Table A1.1 MICS-ECDI2030 Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Can (name) recognize at least 5 letters of the alphabet?	Yes/No	Easy	caregiver
Can (name) write (his/her) name?	Yes/No	Easy	caregiver
Can (name) recognize all numbers from 1 to 5?	Yes/No	Easy	caregiver
If you ask (name) to give you 3 objects, such as 3 stones or 3 beans, does (he/she) give you the correct amount?	Yes/No	Easy	caregiver
Can (name) count 10 objects, for example 10 fingers or 10 blocks, without mistakes?	Yes/No	Easy	caregiver
Can (name) do an activity, such as colouring or playing with building blocks, without repeatedly asking for help or giving up too quickly?	Yes/No	Easy	caregiver
Does (name) ask about familiar people other than parents when they are not there, for example, "Where is Grandma?"?	Yes/No	Easy	caregiver
Does (name) offer to help someone who seems to need help?	Yes/No	Easy	caregiver
Does (name) get along well with other children?	Yes/No	Easy	caregiver
How often does (name) seem to be very sad or depressed? Would you say: daily, weekly, monthly, a few times a year, or never?	Yes/No	Easy	caregiver
Compared with children of the same age, how much does (name) kick, bite, or hit other children or adults? Would you say: not at all, the same or less, more, or a lot more?	Yes/No	Easy	caregiver

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.1.2 AIM-ECD Caregiver Items

Table A1.2. AIM-ECD Caregiver Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
1. Names at least 10 letters	Yes/No	Easy	caregiver
2. Reads four simple words	Yes/No	Easy	caregiver
3. Reads/follows the text in a correct direction from left to right and from top to bottom? (even if they cannot read)	Yes/No	Easy	caregiver
4. Writes at least three letters or some letters in his/her name	Yes/No	Easy	caregiver
5. Writes a simple word	Yes/No	Easy	caregiver
6. Can count from 1 to 10	Yes/No	Easy	caregiver
7. Can count from 1 to 20	Yes/No	Easy	caregiver
8. Knows the difference between tall and short using two animal examples.	Yes/No	Easy	caregiver
9. Knows the difference between heavy and light using two animal examples.	Yes/No	Easy	caregiver
10. Can tell if it is yesterday, today, or tomorrow	Yes/No	Easy	caregiver
11. Knows that a one-digit number is more than another one-digit number (e.g., 4 is more than 2)	Yes/No	Easy	caregiver
12. Pays attention when doing an activity	Yes/No	Easy	caregiver
13. When asked to do several things, remembers all the instructions	Yes/No	Easy	caregiver
14. S/he is able to plan ahead	Yes/No	Easy	caregiver
15. Stops an activity when told to do so	Yes/No	Easy	caregiver
16. Keeps working at something until s/he is finished	Yes/No	Easy	caregiver
17. Gets along with other children s/he plays with	Yes/No	Easy	caregiver
18. Adjusts easily to transitions (for example, to a new teacher or classroom)	Yes/No	Easy	caregiver
19. Accepts responsibility for his/her actions	Yes/No	Easy	caregiver

Table A1.2. AIM-ECD Caregiver Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
20. Settles down after periods of exciting activity	Yes/No	Easy	caregiver

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.1.3   AIM-ECD Direct Assessment Items

Table A1.3 AIM-ECD Direct Assessment Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Letter identification task (17 items)	Correct/Incorrect	Medium	child
Listening comprehension task (4 items)	Correct/Incorrect	Easy	child
Initial sound discrimination tasks (3 items)	Correct/Incorrect	Medium	child
Letter sound identification tasks (2 items)	Correct/Incorrect	Medium	child
Writing names (1 items)	Correct/Incorrect	Hard	child
Number comparison task (1 items)	Correct/Incorrect	Easy	child
Number identification task (10 items)	Correct/Incorrect	Hard	child
Producing a set task (3 items)	Correct/Incorrect	Hard	child
Simple addition and subtraction task (5 items)	Correct/Incorrect	Hard	child
Mental transformation task (3 items)	Correct/Incorrect	Hard	child
Naming shapes task (3 items)	Correct/Incorrect	Hard	child
Object spatial position identification task (4 items)	Correct/Incorrect	Hard	child
Head Toes Knees Shoulders task (12 items)	Correct/Incorrect	Hard	child
Pencil tap task (12 items)	Correct/Incorrect	Hard	child

Table A1.3 AIM-ECD Direct Assessment Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Forward and backward digit span task (3 items)	Correct/Incorrect	Hard	child
Emotion identification task (1 items)	Correct/Incorrect	Hard	child

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.1.4　MICS Foundational Learning Skills

Table A1.4 MICS Foundational Learning Skills Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Do you read books at home?	Yes/No	Easy	child
Does someone read to you at home?	Yes/No	Easy	child
Did the child read every word in the practice correctly?	Yes/No	Hard	child
Once the reading is done, ask: (How old is Sam?/	Correct/Incorrect	Hard	child
Here is another question: (Who is older: Sam or Tina?/	Correct/Incorrect	Hard	child

Table A1.4 MICS Foundational Learning Skills Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Results of the child's reading. LAST WORD ATTEMPTED Incorrect or missed words (B) are those marked incorrect while reading plus the difference between the number of the last word in the story (English:72/ French:86/ Spanish:70) and the last word attempted (A). If the child did not try to read the story, record '00' as the last word attempted (A).	count	Hard	child
Results of the child's reading. LAST WORD ATTEMPTED Incorrect or missed words (B) are those marked incorrect while reading plus the difference between the number of the last word in the story (English:72/ French:86/ Spanish:70) and the last word attempted (A). If the child did not try to read the story, record '00' as the last word attempted (A).	count	Hard	child
FL21A. Check FL20(B): Did the child incorrectly read or miss (English:8/French:9/Spanish:8) or more words?	Yes/No	Hard	child
(What class is Moses in?)	Correct/Incorrect	Hard	child
(What did Moses see on the way home?)	Correct/Incorrect	Hard	child
(Why did Moses start crying?)	Correct/Incorrect	Hard	child
(Where did Moses fall?)	Correct/Incorrect	Hard	child
Why was Moses happy?)	Correct/Incorrect	Hard	child
. Check FL21B[A-E]: Did the child answer all questions correctly?	Yes/No	Hard	child

Table A1.4 MICS Foundational Learning Skills Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Did the child read every word in the practice correctly?	Yes/No	Hard	child
Once the reading is done, ask: (How many eggs does John have?/	Correct/Incorrect	Hard	child
Who has more eggs: John or Anne?/	Correct/Incorrect	Hard	child
Results of the child's reading. LAST WORD ATTEMPTED Incorrect or missed words (B) are those marked incorrect while reading plus the difference between the number of the last word in the story (English:61/French:62/Spanish:68) and the last word attempted (A). If the child did not try to read the story, record '00' as the last word attempted (A).	count	Hard	child
Results of the child's reading. TOTAL NUMBER OF WORDS INCORRECT OR MISSED Incorrect or missed words (B) are those marked incorrect while reading plus the difference between the number of the last word in the story (English:61/French:62/Spanish:68) and the last word attempted (A). If the child did not try to read the story, record '00' as the last word attempted (A).	count	Hard	child
Check FL21P(B): Did the child incorrectly read or miss (English:7/French:7/Spanish:7) or more words?	Yes/No	Hard	child
How old is Mary?/	Correct/Incorrect	Hard	child

Table A1.4 MICS Foundational Learning Skills Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Who sent Mary to the market	Correct/Incorrect	Hard	child
What was Mary asked to buy?/	Correct/Incorrect	Hard	child
Why did Mary lose the money?/	Correct/Incorrect	Hard	child
Why was Mary happy?/	Correct/Incorrect	Hard	child
Check FL23: Did the child correctly identify two of the first three numbers (9, 12 and 30)?	Yes/No	Hard	child
7 & 5	Correct/Incorrect	Easy	child
11 & 24	Correct/Incorrect	Easy	child
58 & 49	Correct/Incorrect	Easy	child
65 & 67	Correct/Incorrect	Easy	child
146 & 154	Correct/Incorrect	Easy	child
3 + 2	Correct/Incorrect	Easy	child
8 + 6	Correct/Incorrect	Easy	child
7 + 3	Correct/Incorrect	Easy	child
13 + 6	Correct/Incorrect	Easy	child
12 + 24	Correct/Incorrect	Easy	child
Turn to the first practice sheet for pattern recognition. Say: Here are some numbers. 1, 2, __, and 4. Point to each number and blank space and say: What number goes here?	Correct/Incorrect	Medium	child

Table A1.4 MICS Foundational Learning Skills Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Here are some more numbers. 5, 10, 15 and __. Point to each number and blank space and say: What number goes here?	Correct/Incorrect	Medium	child
5, 6, 7, __	Correct/Incorrect	Medium	child
14, 15, __, 17	Correct/Incorrect	Medium	child
20, __, 40, 50	Correct/Incorrect	Medium	child
2, 4, 6, __	Correct/Incorrect	Medium	child
5, 8, 11, __	Correct/Incorrect	Medium	child

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.1.5   PISA Household Survey Module

Table A1.5 PISA Household Survey Module Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
A view with a room - space and shape - interpret - math item	Multiple Choice	Hard	child
Baby Growth - uncertainty and data - formulate - math item	Multiple Choice	Hard	child
Concrete Path - change and relationshipts - formulate - math item	Keyword	Hard	child
Machu Picchu - quanitity - interpret - math item	Keyword	Hard	child
Machu Picchu - quanitity - employ- math item	Keyword	Hard	child

Table A1.5 PISA Household Survey Module Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Shoe Sizes - change and relationships- employ-math item	Keyword	Hard	child
Rhino - integrate and interpret - public - reading item	Multiple Choice	Hard	child
Rhino - reflect and evaluate - public - reading item	Multiple Choice	Hard	child
Rhino - integrate and interpret - public - reading item	Multiple Choice	Hard	child
Telephone - access and retrieve - public- reading item	Keyword	Hard	child
Children's Futures - access and retrieve - educational - reading item	Keyword	Hard	child
Children's Futures - access and retrieve - educational - reading item	Keyword	Hard	child
About a book - integrate and interpret - personal - reading item	Keyword	Hard	child
About a book - integrate and interpret - personal - reading item	True/False	Hard	child
Job Vacancy - access and retrieve - occupational - reading item	Keyword	Hard	child
Kenya Tourism - reflect and evaluate - educational - reading item	Multiple Choice	Hard	child
Kenya Tourism - access and retrieve - educational - reading item	Multiple Choice	Hard	child

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.1.6   Mini-LAMP Module

Table A1.6 Mini-LAMP Survey Module Item Details.

Item Description	Response Format	Phone Adaptability	Respondent
Circle the total in corresponding passage -literacy item	circle correct choice	Hard	adult respondent
trade workshops question - retrieve info from a passage - literacy item	fill in blank	Hard	adult respondent
Which country produces least - retrieve info from a table - numeracy item	fill in blank	Hard	adult respondent
Which countries produced same- retrieve info from a table - numeracy item	fill in blank	Hard	adult respondent
What was total election - read a passage and understand table - numeracy item	fill in blank	Hard	adult respondent
Calculate total - read from a table- numeracy item	fill in blank	Hard	adult respondent
How much could you buiy - calculate based on a table - numeracy item	fill in blank	Hard	adult respondent
Camel reading passage item 1 - gather infro from passage - literacy item	fill in blank	Hard	adult respondent
Camel reading passage item 2- gather infro from passage - literacy item	fill in blank	Hard	adult respondent
Camel reading passage item 3- gather infro from passage - literacy item	fill in blank	Hard	adult respondent
Read a graph item 1 - retrieve info from graph - numeracy item	fill in blank	Hard	adult respondent
Read a graph item 2 - retrieve info from graph - numeracy item	fill in blank	Hard	adult respondent
Read a graph item 3 - calculate based on info from graph - numeracy item	fill in blank	Hard	adult respondent
Calculate based on table - apply arithmetic - numeracy item	fill in blank	Hard	adult respondent
Read a gauge - calculate using fractions - numeracy item	fill in blank	Hard	adult respondent

Table A1.6 Mini-LAMP Survey Module Item Details.

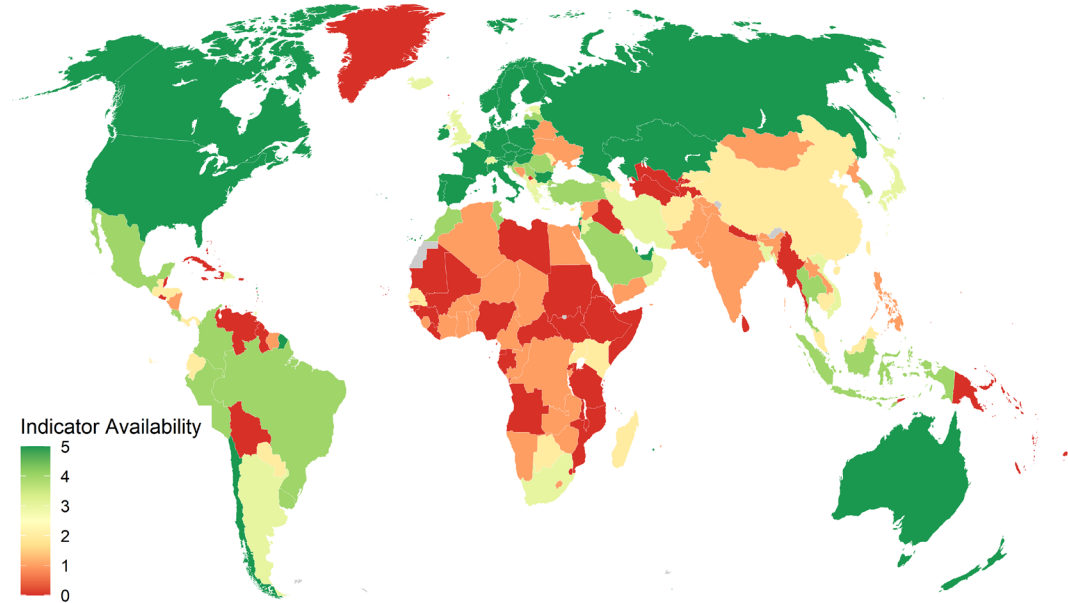| Item Description	Response Format	Phone Adaptability	Respondent

Phone adaptability is based on the authors' judgement and is classified as either easy, medium, or hard 'Easy' items require minimal adaptation for a phone survey. 'Medium' items require some modification for phone use, such as changing item wording, instructions, or some stimulus material may need to be provided by SMS/WhatsApp/etc. 'Hard' items will be very difficult to administer over the phone, as they require stimulus materials such as reading material or require the visual verification of an enumerator.

A.2.1   SDG 4.1.1

4.1.1 Proportion of children and young people achieving a minimum proficiency level in reading and mathematics (%)

Number of Indicators Available 2010-2020

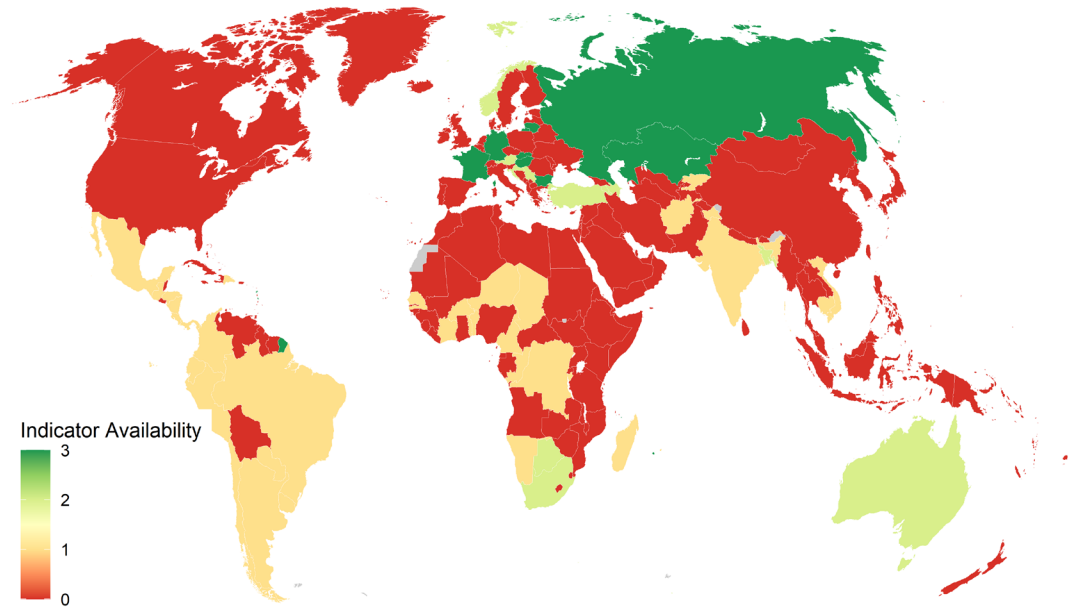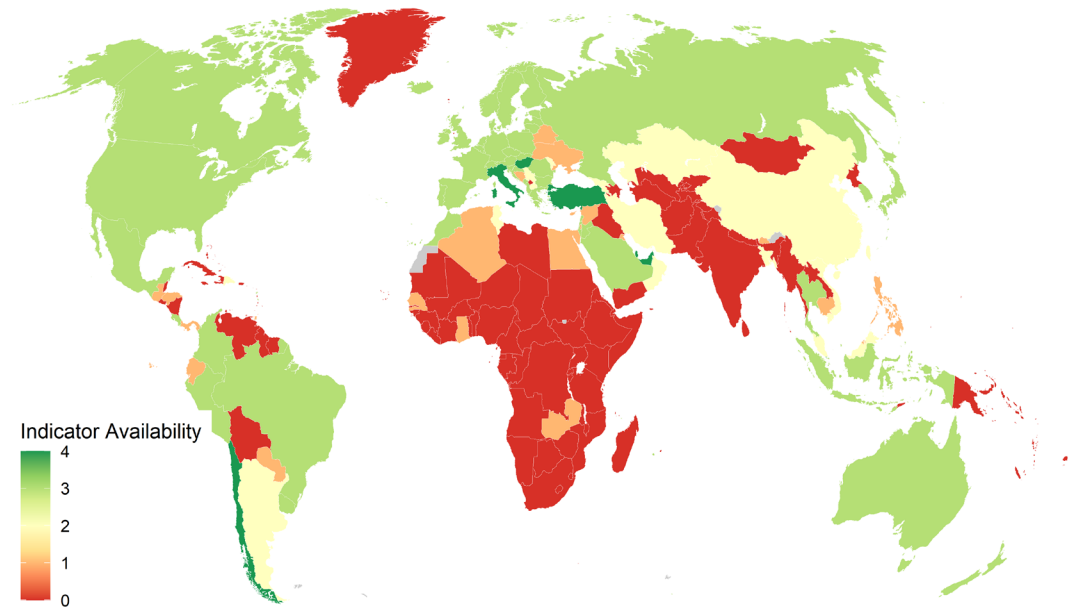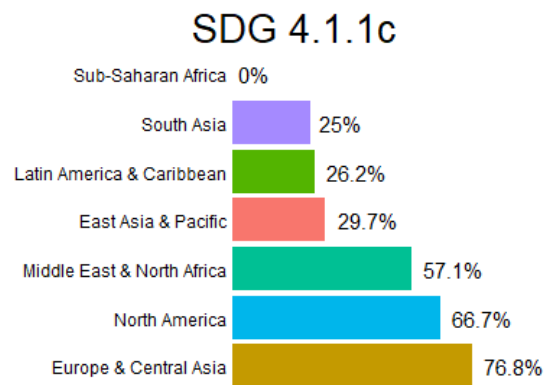Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.1.1a Proportion of children and young people in grade 2/3 achieving a minimum proficiency level in reading and mathematics (%)

Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.1.1b Proportion of children and young people at the end of primary achieving a minimum
proficiency level in reading and mathematics (%)
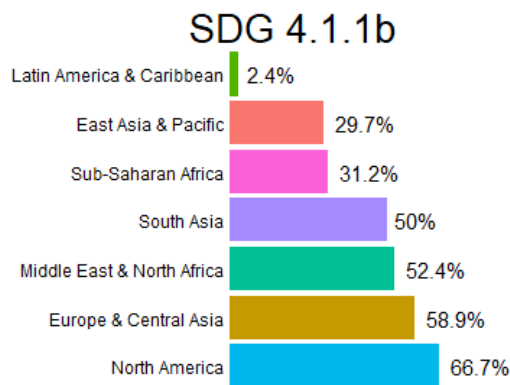
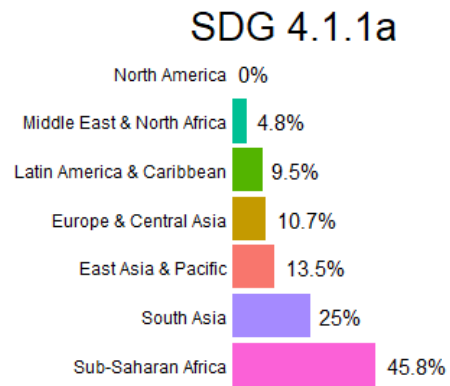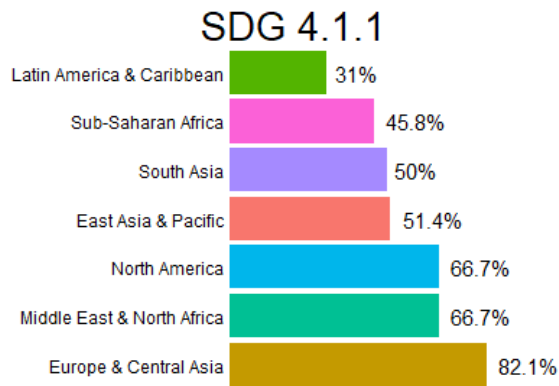Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.1.1c Proportion of children and young people at the end of lower secondary achieving a
minimum proficiency level in reading and mathematics (%)

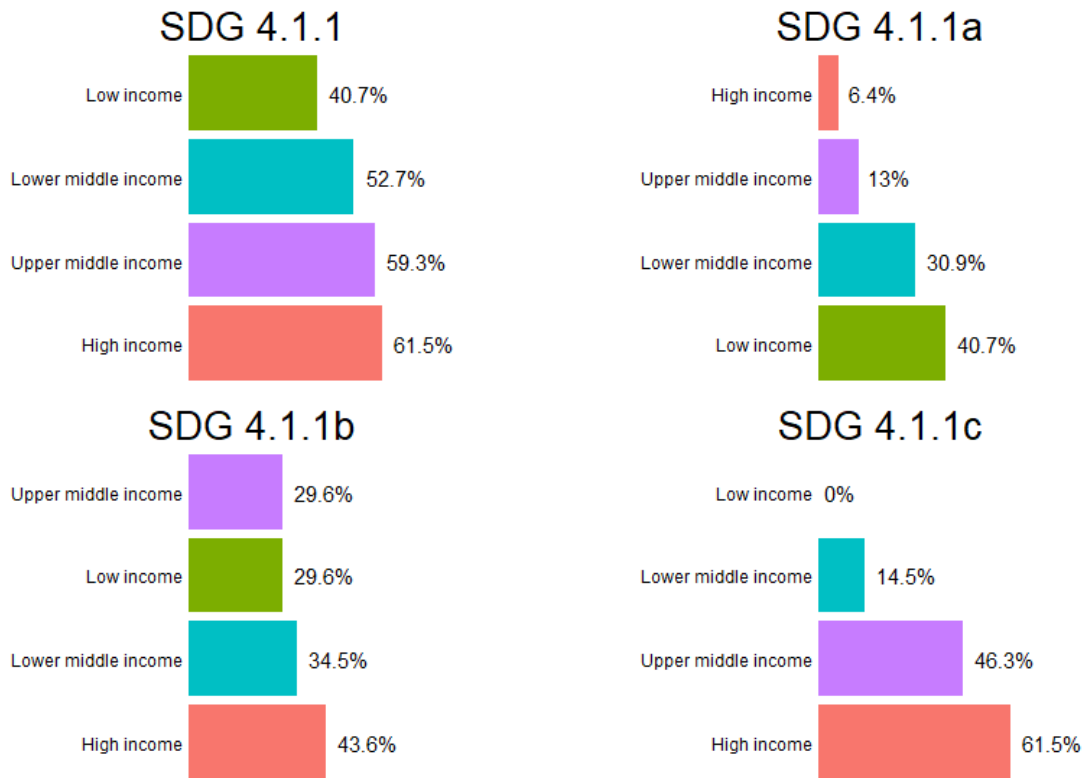Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

Figure A2.1.1 SDG 4.1.1 Indicator Availability by Region

SDG 4.1.1

Region	Percentage
Latin America & Caribbean	31%
Sub-Saharan Africa	45.8%
South Asia	50%
East Asia & Pacific	51.4%
North America	66.7%
Middle East & North Africa	66.7%
Europe & Central Asia	82.1%

SDG 4.1.1a

Region	Percentage
North America	0%
Middle East & North Africa	4.8%
Latin America & Caribbean	9.5%
Europe & Central Asia	10.7%
East Asia & Pacific	13.5%
South Asia	25%
Sub-Saharan Africa	45.8%

SDG 4.1.1b

Region	Percentage
Latin America & Caribbean	2.4%
East Asia & Pacific	29.7%
Sub-Saharan Africa	31.2%
South Asia	50%
Middle East & North Africa	52.4%
Europe & Central Asia	58.9%
North America	66.7%

SDG 4.1.1c

Region	Percentage
Sub-Saharan Africa	0%
South Asia	25%
Latin America & Caribbean	26.2%
East Asia & Pacific	29.7%
Middle East & North Africa	57.1%
North America	66.7%
Europe & Central Asia	76.8%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each region with at least one SDG indicator between 2017 and 2020.

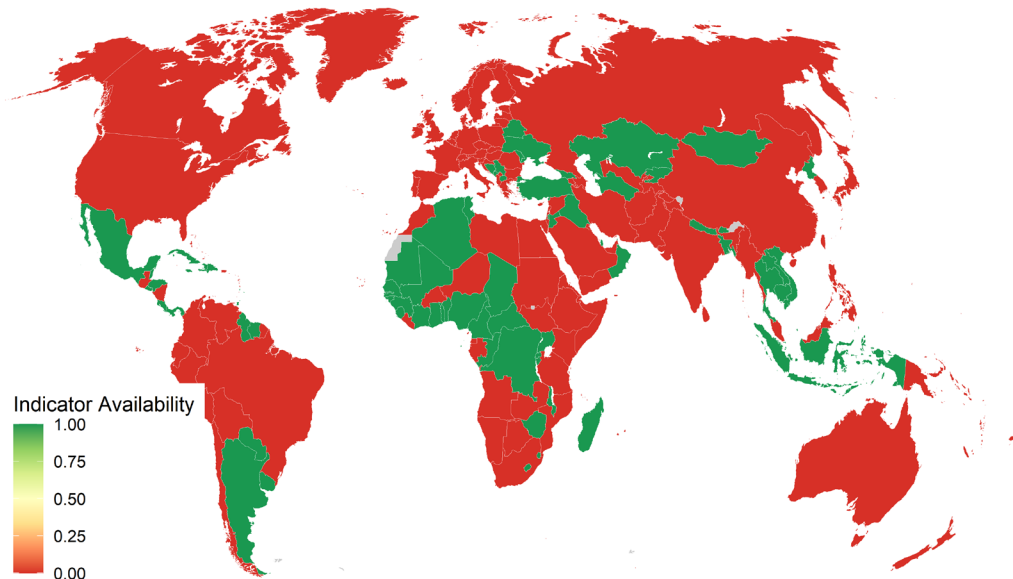Figure A2.1.2 SDG 4.1.1 Indicator Availability by Income

SDG 4.1.1

Low income	40.7%
Lower middle income	52.7%
Upper middle income	59.3%
High income	61.5%

SDG 4.1.1a

High income	6.4%
Upper middle income	13%
Lower middle income	30.9%
Low income	40.7%

SDG 4.1.1b

Upper middle income	29.6%
Low income	29.6%
Lower middle income	34.5%
High income	43.6%

SDG 4.1.1c

Low income	0%
Lower middle income	14.5%
Upper middle income	46.3%
High income	61.5%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each income group with at least one SDG indicator between 2017 and 2020.

A.2.2 SDG 4.2.1

4.2.1 Proportion of children aged 36-59 months who are developmentally on track in at least three of the following domains: literacy-numeracy, physical development, social-emotional development, and learning (% of children aged 36-59 months)

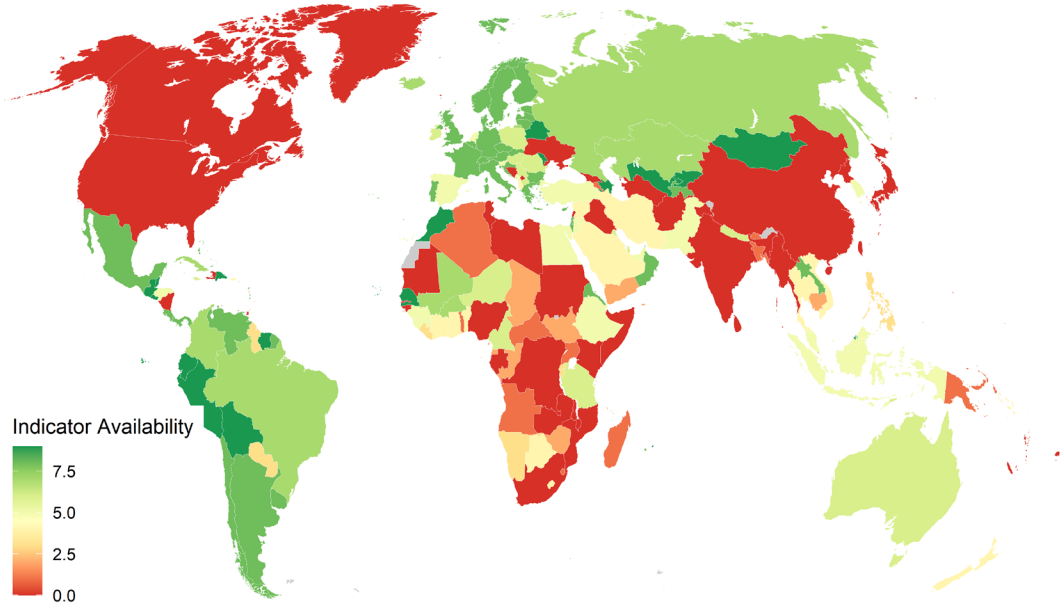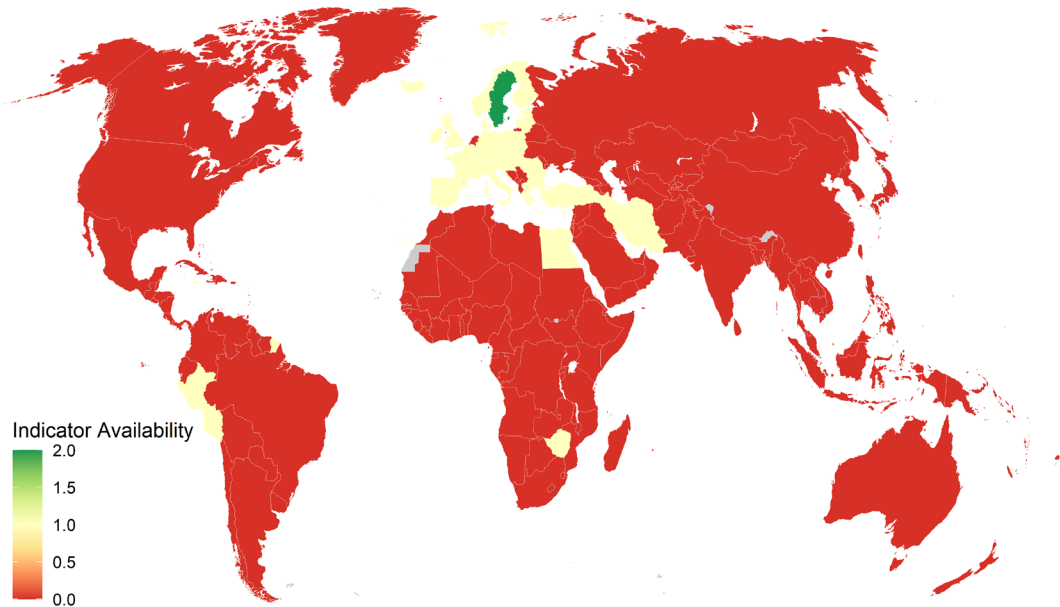Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.5.1 Gender parity index for participation rate in organized learning (one year before the official primary entry age), (ratio)

Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.5.1 Gender parity index for youth/adults with information and communications technology (ICT) skills, by type of skill (ratio)
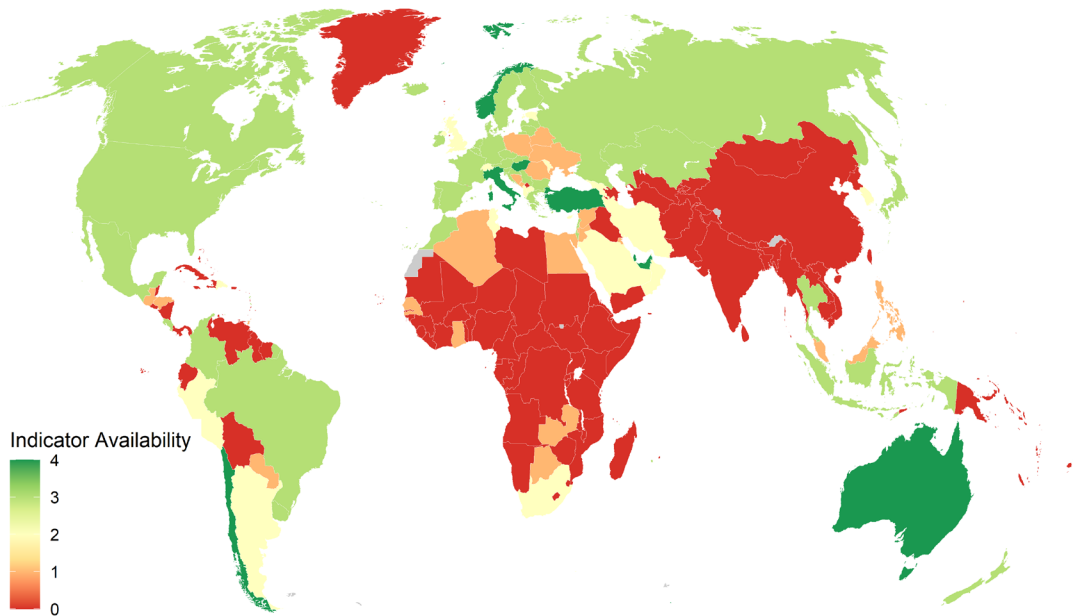
Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

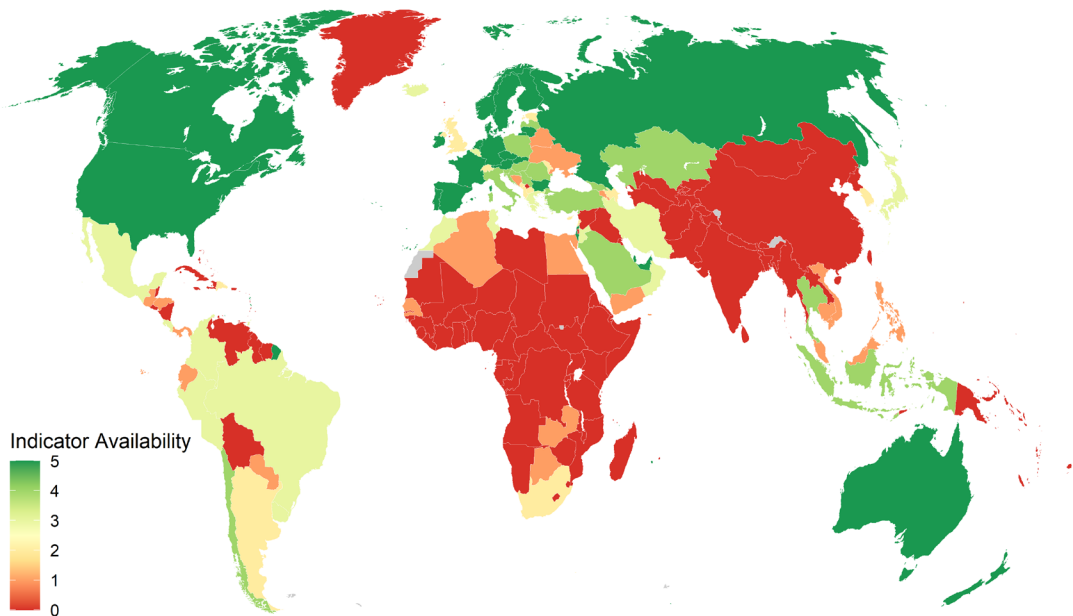## 4.5.1 Native parity index for achievement (ratio)

Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

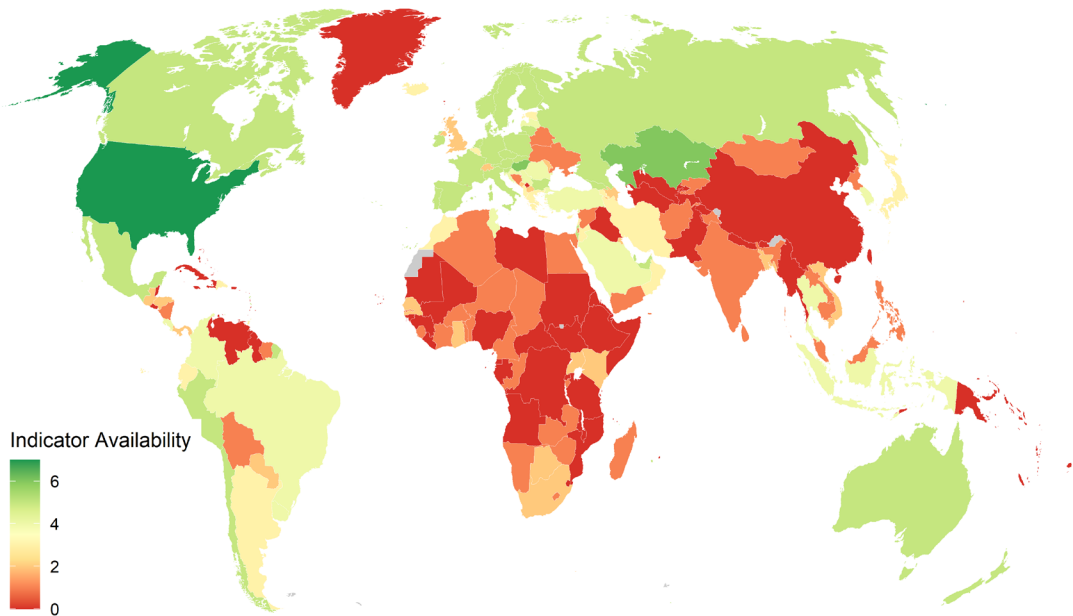## 4.5.1 Language test parity index for achievement (ratio)

Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

4.5.1 Gender parity index for achievement (ratio)
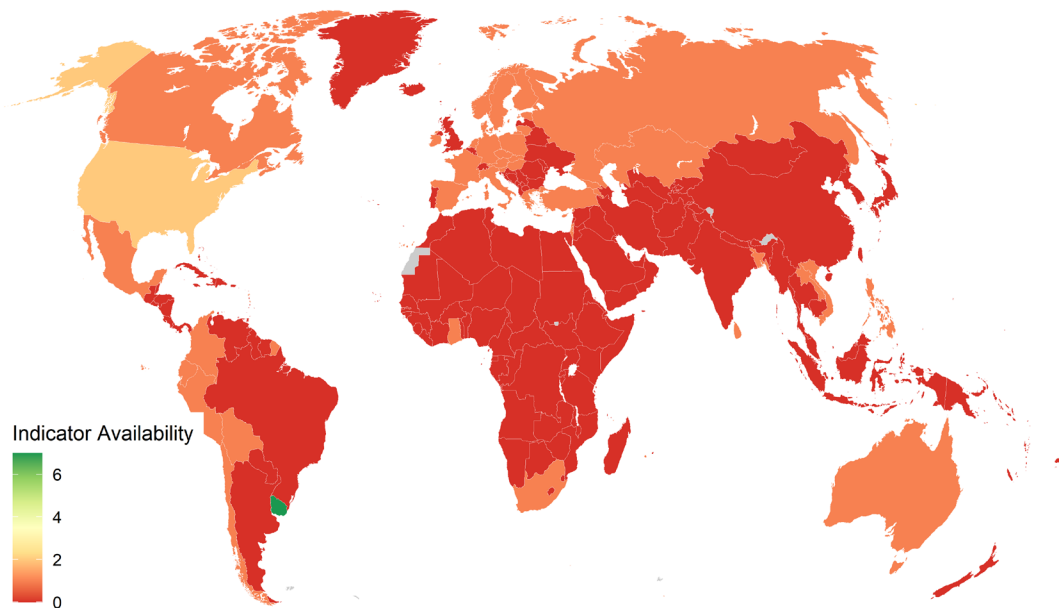Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

A.2.4   SDG 4.6.1

4.6.1 Proportion of population achieving at least a fixed level of proficiency in functional skills, by sex, age and type of skill (%)
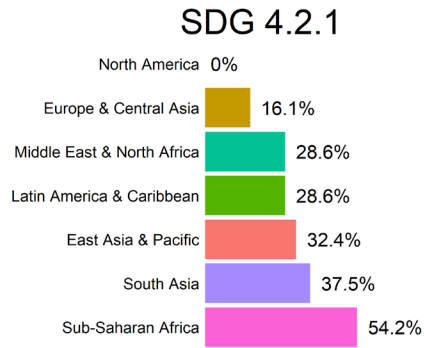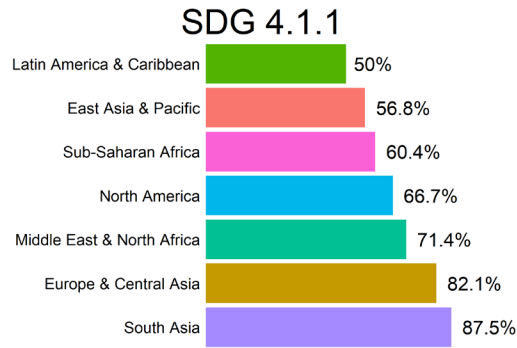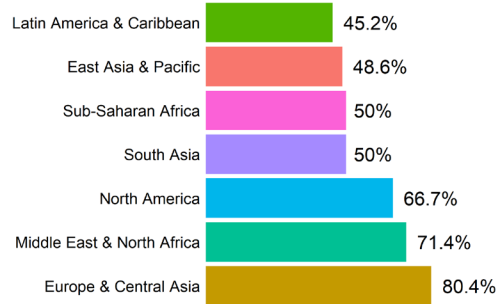Number of Indicators Available 2010-2020

Source: UN Global SDG database. The figures show the availability of SDG indicators by country.

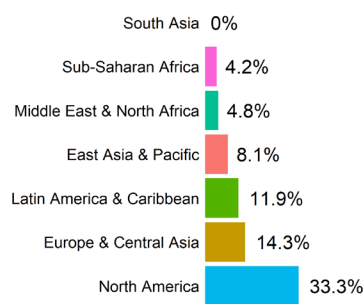## A.2.5   SDG Indicator Availability across 7 years

SDG Indicator Availability by Region across 7 years

51

SDG 4.1.1

Region	Percentage
Latin America & Caribbean	50%
East Asia & Pacific	56.8%
Sub-Saharan Africa	60.4%
North America	66.7%
Middle East & North Africa	71.4%
Europe & Central Asia	82.1%
South Asia	87.5%

SDG 4.2.1

Region	Percentage
North America	0%
Europe & Central Asia	16.1%
Middle East & North Africa	28.6%
Latin America & Caribbean	28.6%
East Asia & Pacific	32.4%
South Asia	37.5%
Sub-Saharan Africa	54.2%

SDG 4.5.1 Gender Learning Parity

Region	Percentage
Latin America & Caribbean	45.2%
East Asia & Pacific	48.6%
Sub-Saharan Africa	50%
South Asia	50%
North America	66.7%
Middle East & North Africa	71.4%
Europe & Central Asia	80.4%

SDG 4.6.1

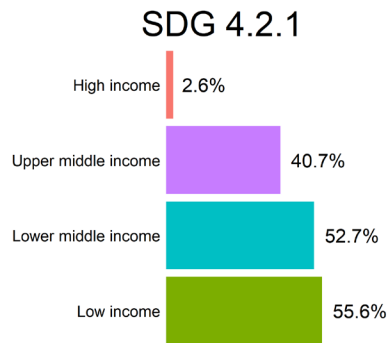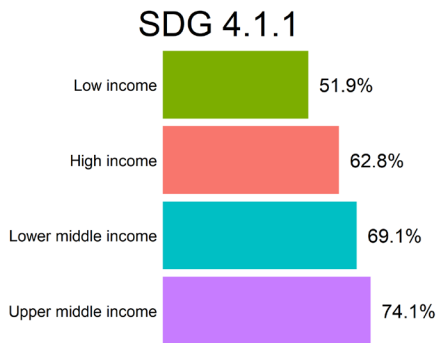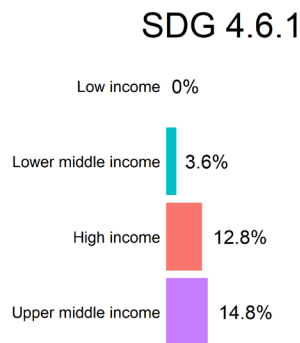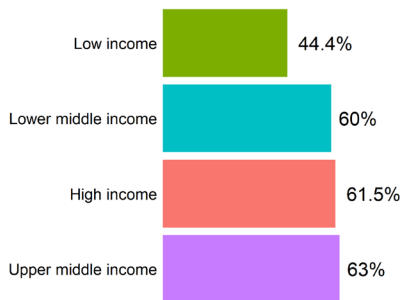| Region	Percentage
South Asia	0%
Sub-Saharan Africa	4.2%
Middle East & North Africa	4.8%
East Asia & Pacific	8.1%
Latin America & Caribbean	11.9%
Europe & Central Asia	14.3%
North America	33.3%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each region with at least one SDG indicator between 2013 and 2020.

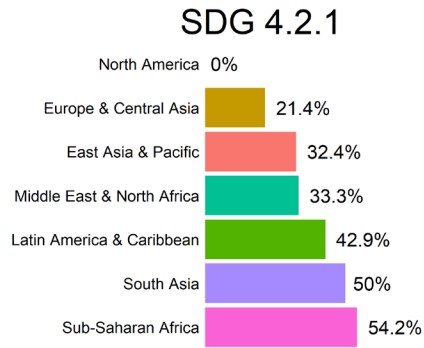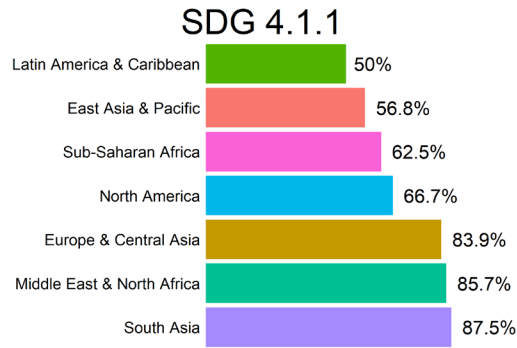SDG Indicator Availability by Income Level across 7 years

SDG 4.1.1

Income Level	Percentage
Low income	51.9%
High income	62.8%
Lower middle income	69.1%
Upper middle income	74.1%

SDG 4.2.1

Income Level	Percentage
High income	2.6%
Upper middle income	40.7%
Lower middle income	52.7%
Low income	55.6%

SDG 4.5.1 Gender Learning Parity

Income Level	Percentage
Low income	44.4%
Lower middle income	60%
High income	61.5%
Upper middle income	63%

SDG 4.6.1

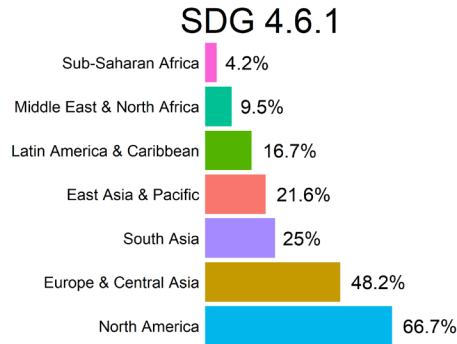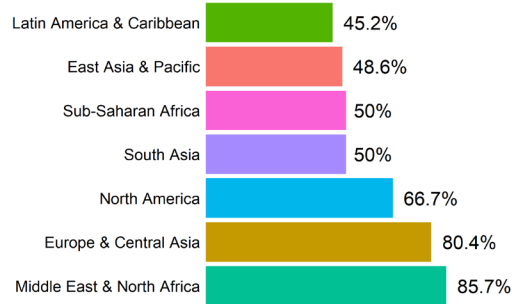| Income Level	Percentage
Low income	0%
Lower middle income	3.6%
High income	12.8%
Upper middle income	14.8%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each income group with at least one SDG indicator between 2013 and 2020.

A.2.5   SDG Indicator Availability across 15 years

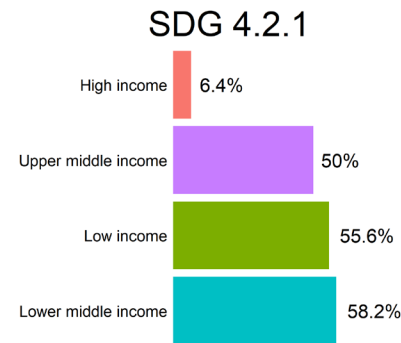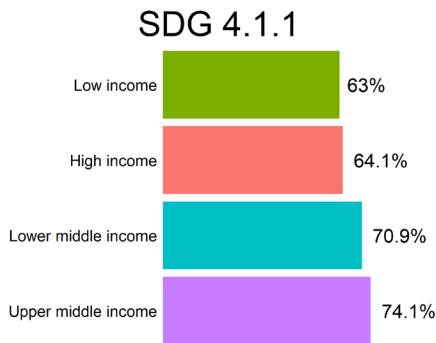SDG Indicator Availability by Region across 15 years

SDG 4.1.1

Region	%
Latin America & Caribbean	50%
East Asia & Pacific	56.8%
Sub-Saharan Africa	62.5%
North America	66.7%
Europe & Central Asia	83.9%
Middle East & North Africa	85.7%
South Asia	87.5%

SDG 4.2.1

Region	%
North America	0%
Europe & Central Asia	21.4%
East Asia & Pacific	32.4%
Middle East & North Africa	33.3%
Latin America & Caribbean	42.9%
South Asia	50%
Sub-Saharan Africa	54.2%

SDG 4.5.1 Gender Learning Parity

Region	%
Latin America & Caribbean	45.2%
East Asia & Pacific	48.6%
Sub-Saharan Africa	50%
South Asia	50%
North America	66.7%
Europe & Central Asia	80.4%
Middle East & North Africa	85.7%

SDG 4.6.1

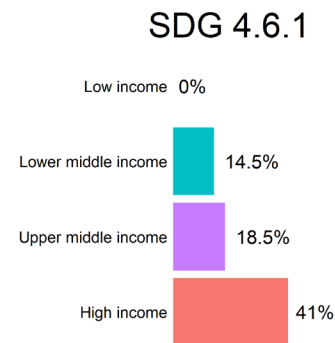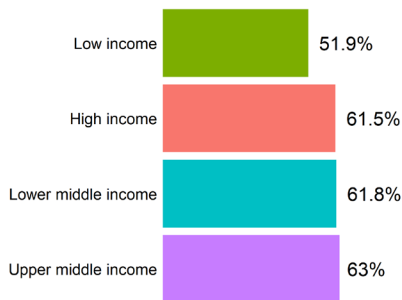| Region	%
Sub-Saharan Africa	4.2%
Middle East & North Africa	9.5%
Latin America & Caribbean	16.7%
East Asia & Pacific	21.6%
South Asia	25%
Europe & Central Asia	48.2%
North America	66.7%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each region with at least one SDG indicator between 2005 and 2020.

SDG Indicator Availability by Income Level across 15 years

SDG 4.1.1

Income Level	%
Low income	63%
High income	64.1%
Lower middle income	70.9%
Upper middle income	74.1%

SDG 4.2.1

Income Level	%
High income	6.4%
Upper middle income	50%
Low income	55.6%
Lower middle income	58.2%

SDG 4.5.1 Gender Learning Parity

Income Level	%
Low income	51.9%
High income	61.5%
Lower middle income	61.8%
Upper middle income	63%

SDG 4.6.1

Income Level	%
Low income	0%
Lower middle income	14.5%
Upper middle income	18.5%
High income	41%

Source: UN Global SDG Database.
Figure shows the percentage of countries in each income group with at least one SDG indicator between 2005 and 2020.

References

Abay, Kibrom A., Guush Berhane, John F. Hoddinott, and Kibrom Tafere. 2021. "Assessing Response Fatigue in Phone Surveys: Experimental Evidence on Dietary Diversity in Ethiopia." WPS9636. Washington, DC: World Bank. https://doi.org/10.2499/p15738coll2.134381.

ACER-GEM. 2020. "Refinement of Minimum Proficiency Levels." Technical Report. http://tcg.uis.unesco.org/wp-content/uploads/sites/4/2020/10/WG-GAML-6-Minimum-Proficiency-Levels_revised.pdf.

Azevedo, João Pedro, Diana Goldemberg, Silvia Montoya, Reema Nayar, Halsey Rogers, Jaime Saavedra, and Brian William Stacy. 2021. "Will Every Child Be Able to Read by 2030? Defining Learning Poverty and Mapping the Dimensions of the Challenge."

Bjarkefur, K., Cardoso de Andrade, L., Daniels, B., & Jones, M. R. (2021). Development Research in Practice. https://openknowledge.worldbank.org/handle/10986/35594

Brennan, P., & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures. BMJ: British Medical Journal, 304(6840), 1491.

Dabalen, Andrew, Alvin Etang, Johannes Hoogeveen, Elvis Mushi, Youdi Schipper, and Johannes von Engelhardt. 2016. *Mobile Phone Panel Surveys in Developing Countries*. The World Bank. https://EconPapers.repec.org/RePEc:wbk:wbpubs:24595.

Dang, Hai-Anh H, and Umar Serajuddin. 2020. "Tracking the Sustainable Development Goals: Emerging Measurement Challenges and Further Reflections." *World Development* 127: 104570.

Division, United Nations Statistics. 2021. "The Sustainable Development Goals Report 2020."

Filmer, Deon, Margarita Langthaler, Robert Stehrer, and Thomas Vogel. 2018. "Learning to Realize Education's Promise." *World Development Report. The World Bank*.

Gochyyev, P., S. Mizunoya, and M. Cardoso. 2019. "Validity and Reliability of the MICS Foundational Learning Module." 9. New York: UNICEF.

Loizillon, A, N Petrowski, P Britto, and C Cappa. 2017. "Development of the Early Childhood Development Index in MICS Surveys." *New York: UNICEF*.

Manuel Cardoso. 2020. "School-Age Learning Assessment Tools: The Foundational Learning Skills Module." Measuring Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries Workshop. https://pubdocs.worldbank.org/en/836911596531694581/Manuel-SchoolAge.pdf

Mizunoya, Suguru. 2020. "Implementing the Foundational Learning Skills module in MICS". Measuring Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries Workshop. https://thedocs.worldbank.org/en/doc/592281596532791960-0050022020/original/MizunoyaLocalization.pdf

Montoya, Silvia. 2020. "Mini-LAMP: A Cost Effective Way to Assess Literacy and Numeracy." Measuring Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries Workshop. Washington DC. https://pubdocs.worldbank.org/en/840121596532249848/Montoya-MiniLamp.pdf

Mullis, Ina and Martin, Michael. 2015. "Introducing IEA's LaNA for Developing Countries." Mexico City. https://www.iea.nl/sites/default/files/2019-04/GA56_TIMSS_PIRLS_Mullis_Martin.pdf.

Pushparatnam, Adelle, Diego Armando Luna Bazaldua, Alaka Holla, João Pedro Azevedo, Marguerite Clarke, and Amanda Devercelli. 2021. "Measuring Early Childhood Development Among 4–6 Year Olds: The Identification of Psychometrically Robust Items Across Diverse Contexts." *Frontiers in Public Health* 9: 17.

Series, Integrated Public Use Microdata. 2021. "IPUMS-USA."

"Strand C Technical Report - PISA." n.d. Accessed May 6, 2021. https://www.oecd.org/pisa/pisa-for-development/strand-c-technical-report/.

Tamassia, Claudia. 2020. "Adaptive Testing Trade Offs".   Measuring Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries Workshop. https://pubdocs.worldbank.org/en/315031596532810099/Tamassia-Adaptive-Testing.pdf

UIS. 2018. "Mini-LAMP for Monitoring Progress Towards SDG 4.6.1." UNESCO Institute for Statistics.

UNESCO and ACER. 2021. "COVID-19: Monitoring Impacts on Learning Outcomes (MILO)." http://gaml.uis.unesco.org/wp-content/uploads/sites/2/2021/03/MILO-STUDY-DESIGN-v3.1-08.04.21.pdf.

Wild, Michael. 2020. "Tablet (CAPI) vs. Paper (PAPI): do we really have a choice?" Measuring Learning through the Lifecycle in Multi-Topic National Household Surveys: Options for low-income countries Workshop. https://pubdocs.worldbank.org/en/846961596532551146/Wild-Capi-Papi.pdf

Woessmann, Ludger. "The importance of school systems: Evidence from international differences in student achievement." *Journal of Economic Perspectives* 30, no. 3 (2016): 3-32.